



The Stochastic Basis of Somatic Variation

Tiago Matias Machado dos Santos Seara Paixão

Dissertação de Doutoramento em Ciências Biomédicas

2007

Tiago Matias Machado dos Santos Seara Paixão

The Stochastic Basis of Somatic Variation

Dissertação de Candidatura ao grau de Doutor
em Ciências Biomédicas, submetida ao Instituto
de Ciências Biomédicas de Abel Salazar da
Universidade do Porto.

Orientador – Doutor Jorge Carneiro

Categoria – Investigador Principal

Afiliação – Instituto Gulbenkian de Ciência

Laboratório Associado de Oeiras

Co-orientador – Doutor Manuel Vilanova

Categoria – Professor Associado

Afiliação – Instituto de Ciências Biomédicas
Abel Salazar da Universidade do Porto.

Acknowledgments

I would like to acknowledge the financial and logistic support of Fundação para a Ciência e Tecnologia (FCT SFRH/BD/10550/2002) and of the Instituto Gulbenkian de Ciência (IGC). This work would not have been possible without the contributions of many different people at both the personal and scientific level. Too all, I am deeply grateful. Specifically, I would like to thank:

To Jorge Carneiro for taking a “dumb physicist” as his student. Thank you for being a friend and a mentor, for all the patience and for always being there. Thanks for continuously challenging me, for broadening my horizons and showing me that science is also about generosity and friendship. Thanks to Filipa for introducing me to Jorge and for always putting up with barbaric invasions of theoretical immunologists into her house with a smile.

To Manuel Vilanova, for co-supervising this thesis, thus joining this walk through strange territories.

To Nuno Sepúlveda and Rui Gardner for always having time to listen, for sharing all moments of this thesis, my highs and lows, for having welcomed me so well in the group, for being good friends. To all past and present members of the Theoretical Immunology Group, José Faro, João Sousa, Kalet, Tiago Carvalho, Ramiro Magno, Carline van den Dool, Luciana, Daniel, Zé Afonso and Íris Vilarés for helping create such a friendly and stimulating environment.

To my thesis committee at the IGC, Jocelyne Demengeot and Isabel Gordo, for helpful advice and for their patience and understanding.

To all my friends at the IGC, especially to Iris Caramalho and Joana Monteiro for being good friends and having the time to listen.

To my best friend, Daniel Ferreira, for all the long nights talking, for his continuous enthusiasm about everything and for being annoyingly inquisitive to the point of making me doubt myself. Above all, thanks for being a good friend.

To Joana, with whom I shared the better part of my life, for always pushing and encouraging me. For making me what I am today. “Não há tempo nem distância”.

To Lurdes, for keeping me sane while writing this thesis, for all the joy for living, for all the “pitas” at late hours and for being such a good friend when I needed most.

Aos meus pais, por todo o apoio, por terem sempre acreditado em mim e por terem estado sempre presentes apesar da minha tendência para “nunca dar notícias”. Às vezes, o meu trabalho não está primeiro. Aos meus irmãos, por toda a paciência necessária a quem vive com alguém que está a fazer uma tese de doutoramento.

Declaração

Os resultados incluídos nesta Dissertação fazem parte de artigos científicos indicados abaixo, alguns dos quais já publicados em revistas internacionais. Do trabalho realizado durante a tese resultaram também outras publicações que caem fora do âmbito da presente tese e que se encontram também indicados abaixo.

Manuscritos Publicados:

Tiago Paixão, Tiago P Carvalho, Dinis P Calado, Jorge Carneiro, “**Quantitative insights into stochastic monoallelic expression of cytokine genes**”, *Immunology and Cell Biology* (2007) 85: 315-322.

Dinis Pedro Calado, Tiago Paixão, Dan Holmberg and Matthias Haury, “**Stochastic Monoallelic Expression of IL-10 in T Cells**” *The Journal of Immunology*, (2006) 177: 5358–5364.

Jorge Carneiro, Tiago Paixão, Dejan Milutinovic, João Sousa, Kalet Leon, Rui Gardner and José Faro, “**Immunological Self-Tolerance: Lessons from mathematical models**” *Journal of Computational and Applied Mathematics* (2005) 184: 77–100

Manuscritos Inéditos:

Tiago Paixão and Jorge Carneiro, “**The role of protein fluctuation structure on adaptation dynamics**” – *in preparation*

Tiago Paixão and Jorge Carneiro, “**The impact of heterogeneity in common network motifs**” – *in preparation*

Tiago Paixão and Jorge Carneiro, “**Structure and Function of the lognormal distribution in protein expression per cell**” – *in preparation*

Outras Publicações:

Jorge Carneiro, Kalet Leon, Íris Caramalho, Carline van den Dool, Rui Gardner, Vanessa Oliveira, Marie-Louise Bergman, Nuno Sepúlveda, Tiago Paixão, Jose Faro, Jocelyne Demengeot, “**When three is not a crowd: a Crossregulation Model of the dynamics and repertoire selection of regulatory CD4+ T cells**” *Immunological Reviews*, 216 (1): 48-68

Table of Contents

Acknowledgments	iii
Declaração	v
Manuscritos Publicados:	v
Manuscritos Inéditos:	v
Outras Publicações:	v
Summary	ix
Sumário	xi
Sommaire	xiii
Introduction	1
Outline of the Thesis	11
References	13
The origin of the lognormal distribution	19
Structure and function of lognormal distribution in protein expression per cell	21
Introduction	23
Results	25
Many sources of multiplicative noise in metabolically stationary cells	25
Log-normal distribution as an “attractor” of the stochastic biochemical networks in the cell	31
Variance of protein copy number per cell as a selectable quantitative trait	34
Discussion	36
Methods	39
References	39
Stochastic Monoallelic Expression	45
Quantitative insights into stochastic monoallelic expression of cytokine genes	47
Introduction	49
Results	51
Stochastic monoallelic IL-10 gene expression: allele independence, low expression probability and instability at the single cell level	51
Simple stochastic transcription activation predicts predominant monoallelic expression but under unrealistic parameter regimes	53
Stochastic predominant monoallelic expression requires cooperative remodelling of locus transcriptional competence	57
Discussion	61
References	66
Implications for cell signaling and population responses	71

The Impact of Heterogeneity In Common Network Motifs.....	73
Introduction	75
Results.....	76
Hyperbolic responses.....	76
Sigmoidal responses.....	80
Feedback and bifurcation induced switches.....	83
Discussion.....	85
References.....	87
Fluctuation Structure and Adaptation Dynamics	91
The Role of Protein Fluctuation Structure on Adaptation Dynamics.....	93
Introduction	95
Model description	96
Results.....	97
Populations with higher variances have higher variation-selection equilibrium points.....	97
Different steady state distributions entail different adaptation dynamics	100
Discussion	102
References.....	104
Stochastic Cell-Cell Interactions as a Generator of Somatic Variation Within Lymphocytes	107
Immunological Self-Tolerance: Lessons from Mathematical Modeling.....	109
Introduction	111
Modeling tolerance by tuning of activation thresholds of individual T lymphocytes	113
A Minimal Model	114
Results	117
Discussion.....	119
Methods	121
References.....	124
Discussion.....	129
Timescales and the intrinsic-extrinsic noise question.....	133
Inheritance and Memory	135
Control of stochastic effects	136
The lognormal distribution.....	137
The adaptive role of heterogeneity generated by somatic variation.....	139
Perspectives and final remarks	141
References.....	145

Summary

Molecular biology has ascribed most phenotypic variation to genetic variation, the so called genetic reductionism. As such, the manifestation of non-genetic individuality, the fact that distinct isogenic cells behave differently under the same environmental conditions, has been a long standing challenge. The low numbers of molecules present within a cell leads to a high variance in the rate of molecular encounters, and since the cell functions through molecular reactions which depend on such encounters this means that kinetic rates can be interpreted as stochastic variables. This is especially relevant in gene transcription since one of the reagents, the chromosome itself, is frequently present in 1 or 2 copies. In fact, the stochastic nature of gene transcription has been implicated in the fluctuations of protein numbers in single cells. Therefore the stochastic nature of molecular reactions provides a framework to understand the essence of non-genetic, somatic variation of protein copy numbers, and of the functions they carry out in the cell.

In this thesis, we pursue the subject of non-genetic individuality in cell populations derived from such stochastic processes. We started by identifying the mechanisms underlying the ubiquitous lognormal distribution of protein copy numbers in cell populations. We demonstrated that this distribution can result from common multistep biochemical processes that propagate noise in a multiplicative fashion. We then turn our attention to a manifestation of non-genetic individuality at the level of the gene: stochastic monoallelic expression. We mathematically formalize and challenge against quantitative experimental data several proposed mechanisms that aim to explain this phenomenon and, as result, we propose a general model of transcription regulation that relates stochastic transcriptional activation and epigenetic chromatin modifications, which provide another source of somatic variation.

We then seek the implications of this stochastically driven variation for the cell population. We analyze how the fact that protein levels are heterogeneous in isogenic populations affects the response of the population. We show that this fact has drastic implications in the ability of a population to respond to a stimulus and draw several concerns about the identifiability of models of signal transduction. We then show how the structure of fluctuations of protein levels can lead to somatic adaptation, providing a basis for selection of different fluctuation generating

mechanisms. Finally, we show how a stochastic environment can lead to heterogeneity within the cell. We show that this type of mechanisms can be used within a multicellular organism for regulation and homeostasis.

These thesis provides a new way to look at cell populations and lineages. We argued that stochastic effects inherent to the cells metabolism generates non-genetic individuality in an isogenic population and entails a new form of “determinism” based not on the deterministic dynamics of the single cell but on the evolution of a probability distribution of the underlying mechanisms. We showed the implications of this at several levels: at the level of gene expression, signal transduction and population dynamics. We have shown how a population of cells could use this spontaneous heterogeneity to their benefit in dealing with stochastic environments or as a regulation mechanism. Moreover, we argue that this heterogeneity creates a link between populations of isogenic cells and their environment. In fact, our work shows that the environment can have a determinant role in shaping and maintaining this heterogeneity within a cell population.

Sumário

A biologia molecular moderna equaciona essencialmente toda a variação fenotípica a variabilidade genética, o chamado reducionismo genético. Assim, a questão de individualidade não genética, i.e. o facto de células isogénicas distintas exibirem comportamentos diferenciados apesar de estarem nas mesmas condições ambientais, é um desafio não resolvido. No entanto, é sabido há muito tempo que números reduzidos das moléculas dentro de uma célula levam a que a variância na taxa de encontros entre moléculas. Já que toda a maquinaria celular funciona através de reacções químicas que dependem de contacto directo entre moléculas isto leva a que a taxa das reacções químicas dentro da célula possa ser considerada estocástica. Isto é especialmente verdade no mecanismo de transcrição de genes, visto que uma das moléculas intervenientes, o cromossoma, está frequentemente presente em apenas 1 ou 2 cópias. De facto, a natureza estocástica da transcrição genética tem sido apontada como a origem das flutuações de números de proteínas em células individuais e, de facto, populações isogénicas exibem uma grande heterogeneidade no número de cópias de proteínas em cada célula. Parece, então, que a natureza estocástica das reacções químicas fornecem uma base segundo a qual se pode entender a essência da variação não genética de número de cópias de proteínas.

Nesta tese elabora-se sobre o assunto de individualidade não genética derivada de acontecimentos estocásticos em populações celulares. Identificam-se vários mecanismos comuns no metabolismo celular que podem explicar a forma específica da distribuição do número de cópias em populações celulares. Tenta-se, então elucidar uma outra manifestação de individualidade não genética: a expressão estocástica monoalélica de genes. Para isso, formalizam-se matematicamente vários mecanismos propostos que tentam explicar este fenómeno, comparam-se as suas previsões com dados experimentais e, dada a falta de plausibilidade destes, propõe-se um novo mecanismo de regulação de transcrição que relaciona eventos estocásticos ao nível genético com modificações epigenéticas.

Procuram-se, então, as implicações para uma população desta individualidade não genética que advém de eventos estocásticos. Analisa-se como o facto dos níveis de expressão de proteínas ser heterogéneo numa população isogénica afecta a resposta de uma população a um estímulo. Mostra-se que este facto tem

implicações drásticas para a capacidade de uma população responder a estímulos e levantam-se algumas preocupações relativas à identificabilidade de modelos de transdução de sinal. Mostra-se, em seguida, como as flutuações de número de proteínas pode levar a adaptação somática, fornecendo um substrato no qual a selecção natural pode actuar para a evolução dos mecanismos de geração destas flutuações. Finalmente, mostra-se como um ambiente estocástico pode levar a que uma população isogénica adquira heterogeneidade. Assim, mostra-se como essa heterogeneidade pode ser usada num organismo multicellular como um mecanismo de regulação e homeostase.

Este corpo de trabalho pretende demonstrar que os efeitos estocásticos que permeiam a célula requerem uma nova perspectiva na maneira de olhar para uma população de células. Requer também uma redefinição do conceito de determinismo biológico, visto que o conceito que pode ser entendido como determinismo no funcionamento de uma célula é a evolução de uma distribuição de probabilidades. O facto de os níveis de proteínas fluctuarem leva a que populações celulares consigam co-optar este fenótipo para seu benefício, tanto para adaptação a ambientes estocásticos como num ambiente estável. Mais do que isso, argumenta-se que o próprio ambiente tem um papel fundamental na manutenção e organização da heterogeneidade de uma população. Isto mostra que a relação entre uma população e o seu ambiente é não trivial e que se pode dizer que uma população é, de certa maneira, congénita do seu ambiente, visto que adapta a sua estrutura interna ao ambiente.

Sommaire

La biologie moléculaire a attribué la plupart de variation phénotypique à la variation génétique, le prétendu réductionnisme génétique. En tant que tels, la manifestation de l'individualité non-génétique, le fait que les cellules isogènes distinctes se comportent différemment dans les mêmes conditions environnementales, en tant qu'été un défi de longue date. Les bas nombres de molécules actuelles dans une cellule mène à un désaccord élevé dans le taux de rencontre moléculaire, et puisque la cellule fonctionne par les réactions moléculaires qui dépendent d'une telle rencontre que ceci signifie que des taux cinétiques peuvent être interprétés en tant que variables stochastiques. C'est particulièrement approprié dans la transcription de gène depuis un des réactifs, le chromosome elle-même, est fréquemment présent dans 1 ou 2 copies. En fait, la nature stochastique de la transcription de gène a été impliquée dans les fluctuations des nombres de protéine dans les cellules. Par conséquent la nature stochastique des réactions moléculaires fournit un cadre pour comprendre l'essence de la variation non-génétique et somatique des nombres de copie de protéine, et des fonctions ils effectuent dans la cellule. Dans cette thèse, nous poursuivons le sujet de l'individualité non-génétique dans des populations de cellules dérivées de tels processus stochastiques. Nous avons commencé par identifier les mécanismes sous-tendants la distribution lognormal omniprésente des nombres de copie de protéine dans des populations de cellules. Nous avons démontré que cette distribution peut résulter des processus biochimiques multiples communs qui propagent le bruit d'une mode multiplicative. Nous tournons alors notre attention à une manifestation d'individualité non-génétique au niveau du gène : expression monoallélique stochastique. Nous mathématiquement formalisons et définissons contre des données expérimentales quantitatives plusieurs mécanismes proposés qui visent à expliquer ce phénomène et, comme résultat, nous proposons un modèle général du règlement de transcription qui relie l'activation stochastique de transcriptionnel et les modifications épigénétiques de chromatin, qui fournissent une autre source de variation somatique. Nous cherchons alors les implications de ceci variation stochastiquement conduite à la population de cellules. Nous analysons comment le fait que les niveaux de protéine sont hétérogènes dans les populations isogènes affecte la réponse de la population. Nous prouvons que ce fait a des implications énergiques dans la capacité d'une population de répondre à un stimulus et de dessiner plusieurs soucis concernant l'identifiabilité des modèles du transduction de signal. Nous montrons alors comment la structure des fluctuations des niveaux de

protéine peut mener à l'adaptation somatique, fournissant une base pour le choix de la fluctuation différente produisant des mécanismes. En conclusion, nous montrons comment un environnement stochastique peut mener à l'hétérogénéité dans la cellule. Nous prouvons que ce type de mécanismes peut être employé dans une organisation multicellulaire pour le règlement et l'homéostasie. Cette thèse fournit une nouvelle manière de regarder des populations et des lignées de cellules. Nous avons argué du fait que les effets stochastiques inhérents au métabolisme de cellules produisent de l'individualité non-génétique dans une population isogène et nécessitent une nouvelle forme d'"déterminisme" basé pas sur la dynamique déterministe de la cellule mais sur l'évolution d'une distribution de probabilité des mécanismes fondamentaux. Nous avons montré les implications de ceci à plusieurs niveaux : au niveau de l'expression de gène, de la transduction de signal et de la dynamique de population. Nous avons montré comment une population des cellules pourrait employer cette hétérogénéité spontanée à leur avantage en faisant face aux environnements stochastiques ou comme mécanisme réglementaire. D'ailleurs, nous arguons du fait que cette hétérogénéité crée un lien entre les populations des cellules isogènes et leur environnement. En fait, notre travail prouve que l'environnement peut avoir un rôle déterminant en formant et en maintenant cette hétérogénéité dans une population de cellules.



Introduction

Ever since the enunciation of the Modern Synthesis, - i.e. the integration of Charles Darwin's mechanism of evolution by natural selection, Gregor Mendel's theory of inheritance and random genetic mutation as the source of phenotypic variants on which selection can operate (Huxley, 1942) -, that all phenotypic variance is ascribed essentially to genetic variance. In fact, population genetics is based on the concept of a (in many cases, one-to-one) genotype-phenotype mapping that makes genes the unit of selection. This concept, strengthened by the discovery of DNA as the source of heritable material and that the flow of biological information was one way only (the central dogma of molecular biology (Crick, 1970), in which information stored at the DNA level can only be transferred to DNA through replication and to RNA and from RNA to protein), ascribed genes as the absolute determinants of phenotype, giving rise to a age of genetic determinism. Individuality was defined by genetic content and the process of individuation was simply the expression or unfolding of the information there contained. This concept of genetic determinism was so successful that gave rise to the widespread view of the cell as gene networks. This view, taken to its extreme consequences, entails that cells can be reduced to a string of zeros and ones (Thomas, 1973), which represent the current state of the organism's genome, in terms of "active" genes (1's) and inactive genes (0's), and a set of discrete rules to propagate this state in time.

However, such convictions have been challenged and revised, as they always are, as more knowledge about these processes accumulates. In fact, mechanisms such as horizontal gene transfer in bacteria, DNA methylation and other epigenetic effects are changing our views on the nature of individuality, making us realize that genetic factors are not the only phenotypic determinants. These previous mechanisms establish non-standard relationships in the central dogma and lead to a notion of historical individuality, since the state of a cell is now specified not just by its actual genetic information but also by its history. DNA methylation, for example, allows proteins to alter the flow of information from DNA to RNA (by modifications that repress transcription) and make the current state of the cell dependent on its historical development, since these modifications can be inherited (Gruenert and Cozens, 1991; Wallace and Orr-Weaver, 2005; Martin and Zhang, 2007). Such mechanisms have evolved through selection and are believed to provide some adaptive advantage to the organism. As such, these mechanisms are still specified

by its genetic information, although through more convoluted relationships, and hence are inline with a gene-centric and deterministic view of biology.

However, the issue of somatic (non-genetic) individuality has arisen in several bacterial studies (Delbruck, 1945; Lederberg and lino, 1956; Powell, 1958; Maloney and Rotman, 1973; Spudich and Koshland, 1976), in many different contexts, ranging from cellular heterogeneity of galactosidase concentrations (Maloney and Rotman, 1973) to distributed bacteriophage burst sizes (Delbruck, 1945). These findings, although hard to attribute to genetic causes or inhomogeneities of the environment, were largely overlooked. In fact, organisms are faced with an immense source of somatic variation: chance. Biological systems have long been recognized to operate in extremely noisy regimes, due to the low number of molecules present (Schrodinger, 1956). Although the upshot of Schrödinger's argument were the chromosomes, his aperiodic crystals, his physical intuition, supported by his knowledge of statistical mechanics, was absolutely correct. The cell functions through direct interactions between molecules, which means that these need to have direct contact. It is clear that when there are low copy numbers of these molecules these encounters will display a large variance in terms of when they happen and how frequently, particularly since there is recent evidence that the cytoplasm is a crowded environment, more akin to a crowded polymer solution than to aqueous media (Weiss et al., 2004; Guigas et al., 2007). This, in turn, affects the rates of biological reactions that govern the cellular behavior. However, with the notable exception of genetic mutation, the role of stochastic events within cell biology was largely overlooked and randomness was something that was left out of all cell biology theory until recently.

The first suggestion that stochastic events within the cell might be responsible for these manifestations of non-genetic individuality is due to Spudich and Koshland (Spudich and Koshland, 1976). They analyzed the tumbling pattern of isogenic individual bacteria when subjected to temporal gradients and noted the distribution of the time it took for an individual bacterium to return to its normal tumbling pattern. They found that this distribution was broad and could roughly be approximated by a Gaussian distribution. Moreover, by correlating this with the length of each bacterium, itself a correlate of the phase of cell cycle, they show that this individuality is not related to the phase of cell cycle. Then, by tethering bacteria to surfaces and repeatedly administering chemoattractant and washing it out, they showed that the individual bacterium maintained their sensitivity to the chemoattractant. They also

made the point that poissonian fluctuations of relevant molecules might be responsible for this variation in all these phenomena. This would imply that the total amount of the relevant molecules had to be very small indeed, as the standard deviation of the Poisson distribution grows with the square root of the mean, becoming less and less significant as the number of copies increases. Of course, further developments in this line of thought were hindered due to the difficulty of available single cell techniques. It is notable that Spudich *et al* obtained these results by microscopy and observing just 22 cells.

As stated before, further investigations into the nature and function of this somatic individuality were hindered by a lack of tools, both technical and theoretical, and undoubtedly by the dominant deterministic view pervading biology. Although the stochastic nature of chemical reactions had been known and worked within the framework of statistical mechanics (Gillespie, 1977), mainly through the work of Gillespie and others, the transposition of such knowledge to the process of transcription had not been done until the 90's. The recent reappraisal of these matters was largely due to the trend setting work of McAdams and Arkin (McAdams and Arkin, 1997), who made stochastic simulations of known mechanisms of genetic regulation in prokaryote organisms concluding that these stochastic events could lead to significant heterogeneity at the population level. This led to the development of one of the first stochastic models of gene expression and soon after to the application of the formalism to the stochastic development of lambda phage infected *Escherichia Coli*. In their work (Arkin et al., 1998), the authors assessed the known mechanism of lysis/lysogeny in this organism and investigated the implications of stochastic fluctuations in this system, correctly predicting the fractions of cells on each of the two developmental branches. Their analysis was solely based on the fluctuations due to stochastic nature of gene expression and hence on poissonian fluctuations of intervening molecules. This showed that heterogeneity within a population could be achieved by fluctuations of proteins that originated from the stochastic events related to transcription. Hence, that somatic variation of protein levels could lead heterogeneity of the population.

Since it became obvious that fluctuations in the number of molecules could have an important impact on the outcome of complex networks of interactions, much interest was drawn to the field, sprouting a great deal of both theoretical and experimental works. Although almost all of these works were based on the assumption that only in conditions of low number of molecules there could be an impact of stochastic effects,

it was realized that, due to the complex network of interactions within a cell, these might have a global impact. In fact, Swain *et al* proposed a distinction between intrinsic and extrinsic noise (Swain et al., 2002). In this framework, intrinsic noise would mean all the variability directly attributable to the fluctuations of the rate of transcription and translation of a given gene, while extrinsic noise would mean any other source of variability that contributed the variability of the gene in question. They proposed that, since extrinsic effects would be the same for both alleles of a given gene in the same cell, a dual reporter assay (in which it would be possible to distinguish between the two alleles) would enable a quantification of the two effects. This terminology pervaded the literature but, as the extrinsic noise was loosely defined, a clear identification of the nature of these components remains elusive. One question that remains to be answered is what is the nature of the extrinsic variability and how does it relate (or influence) the intrinsic?

As concepts such as probability of expression, fluctuation rates and noise became common ground in the literature, the prevailing opinion was that these stochastic effects should be in some way controlled. Indeed, even the term that was coined for these effects, “noise”, evokes the detrimental feeling towards such effects, undoubting due to the terminology used in the central dogma, “information flow”, that pervaded and influenced biology ever since. As such, much work has been done on how the cells circuitry could cope with such erratic processes and on how such noise would spread on the cells genetoc and metabolic networks (Ozbudak et al., 2002; Eissing et al., 2005; Pedraza and van Oudenaarden, 2005; Chen and Wang, 2006; Doncic et al., 2006; El-Samad and Khammash, 2006).

Although the somatic variation stemming from stochastic events in the cell's metabolism might appear as detrimental it is easy to spell out some arguments about potential benefits it might confer to unicellular organisms. It is easy to conceive that if a population of clonal somatically heterogeneous bacteria were to explore its environment, it would have advantage if each bacterium had a different sensitivity to a chemoattractant, as that makes for a better dispersal strategy. Some bacteria would be able to follow very steep gradient but missing local minima, while others would be able to make more refined searches. Another example would be when faced with a toxic chemical; some bacteria would be insensitive to it, due to the population's intrinsic heterogeneity and probably at the expense of having a less efficient metabolic pathway. These would then be able to reconstitute the whole population in all their phenotypic variants and carry the genotype to future

generations (Lederberg and lino, 1956; Soll and Kraft, 1988; Balaban et al., 2004). An example of this is the phenomenon of phenotypic switching which is present in several unicellular organisms. In isogenic colonies of these organisms, a number of phenotypic variants coexist and some functions have been ascribed to these. For example, in bacterial populations subjected to ampicillin it was found (Bigger, 1944) that after a few hours of exposure the initial exponential decay of the population changes to a much less pronounced decay curve. This was found to be due to the presence of a preexistent phenotypic variant that is able, after removal of the antibiotic, to generate an equally susceptible population. This could not be ascribed to genetic mutation and was termed bacterial persistence, in opposition to resistance since this phenotype did not correspond to a resistant genetic mutant. Stochastic switching between phenotypes was implied in this phenomenon but the clear mechanism of operation of that switch was not elucidated. In any case, these phenotypic variants seem to have an adaptive value by maintaining a pool of individuals that are able to survive in certain adverse condition at the expense of lower individual growth rate.

In fact, several works (Norris and Baumber, 1999; Balaban et al., 2004; Kussell and Leibler, 2005) began to hypothesize that this phenotypic variation driven by stochastic events might have an adaptive role, by allowing populations of individuals to acquire heterogeneous phenotypes, and hence raising the capability of bet-hedging, i.e. to be able to sustain environmental changes by diversification of phenotypes. Many of these models focused on abstract mechanisms that allowed a single cell to switch between a number of alternative phenotypes, that conferred different selective advantages in different environments. As such, these populations could have an advantage in fluctuating environments, by always having representatives of several phenotypes that would differentially grow better in different environmental conditions. However, most of these works did not directly specify what stochastic events within the cell lead to this spontaneous phenotypic diversification. A few questions remain unanswered: What kind of mechanisms could lead to this phenomenon? Do stochastic fluctuations of some proteins lead to this directly, in the sense that low probability of expression of many genes induces this phenotypic diversification, or is it possible that other mechanisms are in place?

Nevertheless, the fact that this spontaneous diversification seems beneficial for populations of individuals leads to the possibility of selection of its mechanism of generation. In fact, in all these examples there's a shift in point of view from the

individual to the population as a whole. In this perspective the population takes advantage of its naturally occurring diversity to solve problems that would otherwise be intractable by the single cell. This evokes group selection, an evolutionary theory hypothesis that posited that under certain conditions a group could be the unit of selection, with the alleles being fixed on a population due to the benefits that they bestow on the group and regardless of the individual fitness. This idea was largely discredited (Maynard Smith, 1964), as all selection would have to be explained at the level of the individual, not of the group. This is again inline with a gene centric view of biology and evolution. This view has been contested by several voices, the most prominent being Leo Buss in his seminal work "The Evolution of Individuality". In this work Buss argues that the selective forces that gave rise to the multicellular individual were not operating at the individual level but at the level of cell lineages. Selection would act eliminating lineages that would not preserve the multicellular individual, clearly in same spirit as group selection. In fact, he claims that during the history of Life, selection has shifted through multiple levels of selection. As such, recent reappraisals of the concept of group selection have been made and it has reborn under the name of multilevel selection (Bijma et al., 2007). In fact, it was recognized that the interactions between individuals could indirectly affect the reproductive success of the group, creating an abstract environment where the individual was selected. This is even more so in the case of somatic variation, given that we are dealing with a group of one single genotype, and all variability is non-genetic. In fact, the major criticism towards the idea of group selection was that the genetic variance between individuals was much higher than between groups. Hence, although this phenotype under selection is (the phenotypic diversification) manifested at the population level, it is a property of a single genome, making competition between different groups or colonies exhibiting different mechanisms of diversification a possible driving force for selection.

It has been suggested that organisms with increased genetic complexity, such as the eukaryotes, would be less susceptible to stochastic events due to their increased capabilities of genetic regulation (Spudich and Koshland, 1976). This takes the point of view that prokaryotes, not being able to regulate multiple differentiation states due to less regulatory potential, made use of mechanisms that are always present and active and relied on stochasticity of protein levels to make each cell be in only one of those "states". Eukaryotic organisms, on the other hand, being able to sustain more states of cell differentiation, did not have to rely on such "blind" prediction of future circumstances and could rely on environmental cues to make appropriate choices.

In fact, this is akin to phenotypic plasticity that has been defined as the ability of a specific genotype to exhibit alternative traits in response to environmental conditions (biotic or abiotic) (Fordyce, 2006), and has been known for some time, mostly in ecological literature (Agrawal, 2001). This definition, however, refers more to this concept of extra regulatory potential than to non-genetic variability since several aspects of “regular” functioning of many organisms involve such type of response without necessarily involving stochastic events. For example, budding yeast (*Saccharomyces cerevisiae*) are able to choose when to undergo asexual reproduction and when to engage in sexual reproduction based on the presence or absence of a pheromone in the environment, and multicellular organisms display a multitude of cell types that stem from a progenitor in response to specific developmental cues or genetic program.

These are traits that exactly make use of the extra layers of regulation available to eukaryotes. Increased genetic potential means these organisms might be able to genetically hardcode many different responses without having to rely on diversification driven by stochastic events.

As such, prokaryotic organisms have been the organisms of choice for studying somatic variation driven by stochastic events, both for technical and historical reasons but also because it was assumed that these effects were more evident in prokaryotes. This assumption stemmed from many reasons. First, as stated above, prokaryotes lack many regulatory mechanisms that are present in eukaryotes and so it seemed that these effects would have more impact in these more primitive organisms. Second, prokaryotes are typically many times smaller than their eukaryotic counterparts and so their protein copy numbers were vastly reduced compared to the one in eukaryotic organisms. Since low copy number has been seen as the major source for stochasticity within the cell, eukaryotes seemed less prone to stochastic effects. As such, advances regarding these matters in eukaryotic cells were overlooked. However, with the advent of flow cytometry and fluorescence activated cell sorting (FACS) technology (Herzenberg and Sweet, 1976) allowed a deep exploration of the constituents of complex systems (in the sense that are constituted by many interacting populations) such as the immune system. Furthermore, they allowed a quantification of the heterogeneity in protein copy number per cell of eukaryotic, as well as prokaryotic, populations. And indeed these populations showed remarkably broad distributions of these quantities. Moreover, these distributions typically displayed a characteristic shape: the lognormal distribution. This shape cannot be explained by the so-called intrinsic noise and its

generative mechanism is not fully understood. A more complete understanding of the origin of this shape might shed some light into the origin of the aforementioned extrinsic noise, and its relation to stochastic gene expression.

The fact that eukaryotic organisms also display high levels of heterogeneity in protein copy numbers leads to the interesting question of whether stochastic effects play any role in eukaryotes, especially given the rationale above about their ability to control them. In fact, since these organisms are in general bigger and express more copies of each protein and have increased layers of regulation of gene expression it is conceivable that evolution could have been able to constrain stochastic effects within cells, at least at the transcriptional level. However, it was found that transcription in eukaryotes also occurred “randomly and infrequently” (Ross et al., 1994). In fact, the extra mechanisms present in these organisms, allowed for new manifestations of stochastic driven somatic variation. Polyploidy, usually only present in eukaryotic organisms, was one of the features of eukaryotic organism that was deemed to circumvent the inherent stochasticity of transcription (Cook et al., 1998). However, it was found that an increasing number of genes were expressed in a monoallelic fashion (Hollander et al., 1998; Riviere et al., 1998; Nutt and Busslinger, 1999; Calado et al., 2006), and the choice of which allele was to be expressed seemed to be stochastic. So, it seems that at least in some cases, eukaryotic organisms are poised towards noisier gene expression. Could this serve any function? What is the mechanism that prevents the default biallelic expression of a gene?

These observations suggest that somatic variation due to stochastic events is present in all organisms. Particularly in multicellular organisms, it seems that there is no reason why these organisms would not be able to take advantage of an available source of variation: no evolutionary arguments prevent that, since non-genetic variation within the somatic cells is irrelevant for the genome’s perpetuation. In fact, many examples of this exist, the immune system being the most paradigmatic example, as we’ll see below.

The immune system is constituted by several interacting populations and is typically divided into innate and the adaptive immune system. Its main function is to recognize and prevent pathogens from invading the organism. Only vertebrates have developed the adaptive immune system and the term “immune system” is usually reserved to the adaptive immune system. The innate component of the immune system is evolutionarily more ancient and is present in almost all of the multicellular organisms. It performs the above function using a variety of mechanisms but one of

the most common (found on both vertebrates and invertebrates) is using germline encoded receptors that respond to common molecular patterns found on pathogens or other microbes (Pathogen Associated Molecular Patterns - PAMPs). In fact, innate immunity is universal in the fact that recognition is achieved by mechanisms hardwired in the organism's genome by evolution, and is essentially blind to all other patterns not present in its finite repertoire. The adaptive immune system is constituted by lymphocytes which are further classified in B, T, NK and dendritic cells, all derived from the common lymphoid progenitor (CLP) (Kondo et al., 1997). The B and T cell type precursors undergo their development in the bone marrow and thymus (respectively) where their TCR and BCR genes suffer a process of somatic recombination to form functional antigen receptors. These receptors are the core of the adaptive immune system as each is virtually unique, due to the stochastic process occurring in their rearrangement. In the case of T cells, these lymphocytes undergo a process of selection in order to guarantee that their receptor genes have productive rearrangements (functional TCRs) and do not have potentially harmful specificities, i.e. self-reactive specificities.

It is interesting to note the parallels between what we called "genetic individuality" and the innate immune system and the emergence of somatic variation in both cases, in the form of the adaptive immune system and the stochastic driven non-genetic variation. It seems that both instances of somatic variation are a way of an organism to predict the unpredictable, by making use of large number of cells and random variation.

Furthermore, the immune system relies on interactions between populations for its proper functioning. This makes it a good source of examples for studying how this somatic variation propagates to interaction between populations and the possibility of a multicellular organism taking advantage of that.

Outline of the Thesis

The present thesis aims to explore several questions regarding the origin, implications and putative function of non-genetic individuality brought by stochastic events, drawing on several aspects of the immune system for concrete examples. We focus mainly on variability at the level of protein copy number and its implications in cell populations but we also explore other aspects of somatic variation brought by stochastic events. This thesis is organized in 5 chapters, excluding the present introduction and general discussion, which are presented in the form of scientific

publications (already published, submitted or in final stages of preparation) with specific introductions and discussions.

In chapter II, we discuss the origins of eukaryotic heterogeneity in protein copy number and elaborate on possible models that generate a characteristic shape of this distribution and critically examine other works that propose different explanations for this particular shape. We also show that this particular shape is extremely versatile in reproducing population response curves when individual cells are responding in an all-or-none fashion.

In chapter III, we elaborate on the mechanism by which stochastic monoallelic expression of cytokine genes occurs and its biological function. We formalize two different mechanisms; one based on simple stochastic gene expression and the other based on chromatin modifications. Taking advantage of experimental data available from a mouse model of allelic expression in which the two alleles of the IL-10 gene could be distinguished we were able to exclude the first former model. We conclude that, in order for the model based on epigenetic modifications to be able to explain the data, several intermediate states were necessary and that some kind of bistability, which we propose to be mediated by feedback loops, is needed to be in operation in these states. From these results we speculate on the adaptive value of this specific kind of non-genetic variation. We also take the opportunity to discuss the issue of intrinsic and extrinsic noise, as similar experimental models were used to make this terminology popular.

In chapter IV, we pursue the question of the implications of population heterogeneity in signal transduction and other metabolic pathways, assuming a static heterogeneity, i.e. that the processes that give rise to such heterogeneity are much slower than the processes we are studying. We characterize how both the dynamics and steady states are affected by variation in relevant components of simple biochemical modules by constructively concatenating and rewiring very simple modules. We explore how heterogeneity in simple modules brings about population effects usually associated with more complex modules, such as bistability. We also explore how the traditional techniques of model identification, usually based on the population averages, are hindered in the presence of heterogeneity in the module components.

In chapter V, we address the question of the adaptive value of non-genetic variation and to what extent the dynamics of variation affects the fitness of a distributed population. In order to do so we develop a mathematical competition model between populations with different fluctuation characteristics. We derive conditions in which

faster fluctuations rates can be an advantage and compare different types of fluctuation distributions, suggesting that these might be a selectable trait.

In chapter VI, we analyze how extracellular stochastic events bring about heterogeneity in activation thresholds of T cells. Specifically, we study how cell-cell contacts modulate concentrations of a kinase and phosphatase that control an intracellular biochemical switch (as studied in chapter IV) which controls proliferation. We show that this extracellular events leads to heterogeneity of activation thresholds and that this heterogeneity offers a mechanism of self-tolerance by prevention of autoreactive T cell expansion.

In the general discussion, we overview several outstanding questions on the literature regarding the matter of stochastic somatic variation. We place these questions in the context of the present thesis and compare our conclusions with related works.

References

- Agrawal, A. A. (2001). "Phenotypic plasticity in the interactions and evolution of species." Science **294**(5541): 321-6.
- Arkin, A., J. Ross and H. H. McAdams (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." Genetics **149**(4): 1633-48.
- Balaban, N. Q., J. Merrin, et al. (2004). "Bacterial persistence as a phenotypic switch." Science **305**(5690): 1622-5.
- Bigger, W. (1944). "Treatment Of Staphylococcal Infections With Penicillin by Intermittent Sterilisation." The Lancet **244**(6320): 497-500.
- Bijma, P., W. M. Muir and J. A. Van Arendonk (2007). "Multilevel selection 1: Quantitative genetics of inheritance and response to selection." Genetics **175**(1): 277-88.
- Calado, D. P., T. Paixao, D. Holmberg and M. Haury (2006). "Stochastic monoallelic expression of IL-10 in T cells." J Immunol **177**(8): 5358-64.
- Chen, B. S. and Y. C. Wang (2006). "On the attenuation and amplification of molecular noise in genetic regulatory networks." BMC Bioinformatics **7**: 52.

- Cook, D. L., A. N. Gerber and S. J. Tapscott (1998). "Modeling stochastic gene expression: implications for haploinsufficiency." Proc Natl Acad Sci U S A **95**(26): 15641-6.
- Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-3.
- Delbruck, M. (1945). "The Burst Size Distribution in the Growth of Bacterial Viruses (Bacteriophages)." J Bacteriol **50**(2): 131-5.
- Doncic, A., E. Ben-Jacob and N. Barkai (2006). "Noise resistance in the spindle assembly checkpoint." Mol Syst Biol **2**: 2006 0027.
- Eissing, T., F. Allgower and E. Bullinger (2005). "Robustness properties of apoptosis models with respect to parameter variations and intrinsic noise." Syst Biol (Stevenage) **152**(4): 221-8.
- El-Samad, H. and M. Khammash (2006). "Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks." Biophys J **90**(10): 3749-61.
- Fordyce, J. A. (2006). "The evolutionary consequences of ecological interactions mediated through phenotypic plasticity." J Exp Biol **209**(Pt 12): 2377-83.
- Gillespie, D. T. (1977). "Exact Stochastic Simulation of Coupled Chemical Reactions." Journal Physical Chemistry **81**: 2340-2361.
- Gruenert, D. C. and A. L. Cozens (1991). "Inheritance of phenotype in mammalian cells: genetic vs. epigenetic mechanisms." Am J Physiol **260**(6 Pt 1): L386-94.
- Guigas, G., C. Kalla and M. Weiss (2007). "Probing the nano-scale viscoelasticity of intracellular fluids in living cells." Biophys J.
- Herzenberg, L. A. and R. G. Sweet (1976). "Fluorescence-activated cell sorting." Sci Am **234**(3): 108-17.
- Hollander, G. A., S. Zuklys, et al. (1998). "Monoallelic expression of the interleukin-2 locus." Science **279**(5359): 2118-21.
- Huxley, J. S. (1942). Evolution: The Modern Synthesis, Allen and Unwin.
- Kondo, M., I. L. Weissman and K. Akashi (1997). "Identification of clonogenic common lymphoid progenitors in mouse bone marrow." Cell **91**(5): 661-72.
- Kussell, E. and S. Leibler (2005). "Phenotypic diversity, population growth, and information in fluctuating environments." Science **309**(5743): 2075-8.
- Lederberg, J. and T. Iino (1956). "Phase Variation in Salmonella." Genetics **41**(5): 743-57.
- Maloney, P. C. and B. Rotman (1973). "Distribution of suboptimally induced -D-galactosidase in Escherichia coli. The enzyme content of individual cells." J Mol Biol **73**(1): 77-91.

- Martin, C. and Y. Zhang (2007). "Mechanisms of epigenetic inheritance." Curr Opin Cell Biol **19**(3): 266-72.
- Maynard Smith, J. (1964). "Group Selection and Kin Selection." Nature **201**: 1145-1147.
- McAdams, H. H. and A. Arkin (1997). "Stochastic mechanisms in gene expression." Proc Natl Acad Sci U S A **94**(3): 814-9.
- Norris, T. L. and A. J. Baumler (1999). "Phase variation of the *lpf* operon is a mechanism to evade cross-immunity between *Salmonella* serotypes." Proc Natl Acad Sci U S A **96**(23): 13393-8.
- Nutt, S. L. and M. Busslinger (1999). "Monoallelic expression of *Pax5*: a paradigm for the haploinsufficiency of mammalian *Pax* genes?" Biol Chem **380**(6): 601-11.
- Ozbudak, E. M., M. Thattai, et al. (2002). "Regulation of noise in the expression of a single gene." Nat Genet **31**(1): 69-73.
- Pedraza, J. M. and A. van Oudenaarden (2005). "Noise propagation in gene networks." Science **307**(5717): 1965-9.
- Powell, E. O. (1958). "An outline of the pattern of bacterial generation times." J Gen Microbiol **18**(2): 382-417.
- Riviere, I., M. J. Sunshine and D. R. Littman (1998). "Regulation of IL-4 expression by activation of individual alleles." Immunity **9**(2): 217-28.
- Ross, I. L., C. M. Browne and D. A. Hume (1994). "Transcription of individual genes in eukaryotic cells occurs randomly and infrequently." Immunol Cell Biol **72**(2): 177-85.
- Schrodinger, E. (1956). What is Life? & Other Scientific Essays, Doubleday Anchor.
- Soll, D. R. and B. Kraft (1988). "A comparison of high frequency switching in the yeast *Candida albicans* and the slime mold *Dictyostelium discoideum*." Dev Genet **9**(4-5): 615-28.
- Spudich, J. L. and D. E. Koshland, Jr. (1976). "Non-genetic individuality: chance in the single cell." Nature **262**(5568): 467-71.
- Swain, P. S., M. B. Elowitz and E. D. Siggia (2002). "Intrinsic and extrinsic contributions to stochasticity in gene expression." Proc Natl Acad Sci U S A **99**(20): 12795-800.
- Thomas, R. (1973). "Boolean formalization of genetic control circuits." J Theor Biol **42**(3): 563-85.
- Wallace, J. A. and T. L. Orr-Weaver (2005). "Replication of heterochromatin: insights into mechanisms of epigenetic inheritance." Chromosoma **114**(6): 389-402.
- Weiss, M., M. Elsner, F. Kartberg and T. Nilsson (2004). "Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells." Biophys J **87**(5): 3518-24.



The origin of the lognormal distribution

Structure and function of lognormal distribution in protein expression per cell

Tiago Paixão* and Jorge Carneiro*

**Instituto Gulbenkian de Ciência*

Summary

In isogenic prokaryotic or eukaryotic cell populations the copy numbers per cell of many constitutive proteins are lognormal distributed. This distribution, which is observed both in quiescent cells and in cells that are growing exponential, is a footprint of the organization and dynamics of the molecular machinery of the cell, and one may gain insights into cell biology by understanding the nature of the stochastic processes that generate it. We show here that common multistep biochemical processes, namely protein biosynthetic chains, biosynthetic or regulatory cascades, and transcription factor aggregation necessary for transcription activation, can, under non trivial conditions, propagate noise in a multiplicative way and thus give rise to products that are lognormal distributed in a cell population. This offers a simple explanation for the ubiquity of this distribution, and predicts that its variance is mainly controlled by the number of steps in the biochemical process regulating the protein expression. We argue that the variance itself is a selectable quantitative trait and substantiate this by deriving a biunivocal mapping between the parameters of a Hill function describing a bulk cell population response and the median and variance of the lognormal distribution of a protein controlling response threshold in individual cells.

Introduction

Proteins are key components of the molecular machinery of the cell, ensuring specific functions that depend on well regulated protein levels. Intriguingly, populations of isogenic prokaryotic or eukaryotic cells show a remarkable heterogeneity in the levels of expression of proteins (McAdams and Arkin, 1997). The ultimate cause for the heterogeneity in protein content is the stochastic effects inherent to the low copy number of the molecules involving in regulating the protein activity (Ozbudak et al., 2002). Low copy number of mRNA molecules and stochastic fluctuations of local concentration due to diffusion are impossible to eliminate, even to the more evolved eukaryotes (McAdams and Arkin, 1999), which have additional layers of genetic regulation (Blake et al., 2003). The nature of this randomness motivated several studies of the stochastic regulation of gene expression (Thattai and van Oudenaarden, 2001; Ozbudak et al., 2002; Chen and Wang, 2006; Doncic et al., 2006; El-Samad and Khammash, 2006). Hitherto, most models of stochastic kinetics of gene expression, based on either the Gillespie-algorithm or stochastic differential equations, propagate intrinsic noise in an additive way predicting protein copy fluctuations that follow distributions of the Poisson family (Kaern et al., 2005). In Poisson distributions, the average and variance are identical, and in the limit of many molecules the distribution tends to a Gaussian symmetric in linear scale. High throughput proteomic analysis of the distribution of many proteins indicates that although variance is proportional to mean, the proportionality constant is greater than the unit. This rules out the possibility that the observed variance in protein copy numbers is due to intrinsic stochastic fluctuations of the proteins themselves, but could still be compatible with additive propagation of extrinsic noise (Blake et al., 2003). But the observed distributions for many constitutive proteins are very different from those expected from additive noise propagation. Thus, instead of symmetric in the number of molecules per cell, they are symmetric in its logarithm (Figure1a). This raises the question on what kind of stochastic process generates these lognormal distributions? What are available models missing about the noise propagation structure?

Distributions of the protein activity that are symmetric in log scale can emerge from stochastic intercellular interactions as those undergone by T lymphocytes when they form serial conjugates with antigen-presenting cells (Carneiro et al., 2005; Milutinovic et al., 2007). However, this is not a general enough mechanism to explain

the ubiquity of lognormal distribution in prokaryotic and eukaryotic cells. Accordingly, the more often evoked explanation is the multiplicative propagation of noise during exponential growth of the cell populations. Krishna et al. (Krishna et al., 2005) argued that long-tailed distributions would emerge when the average number of protein copies is small and growing exponentially. Similarly, Furusawa et al. (Furusawa et al., 2005) concluded from simulations of catalytic molecular networks that efficient self-reproduction during steady exponential growth entails multiplicative noise propagation and leads log-normally distributed protein copy numbers. Although these two mechanistic hypotheses differ in their specific details, they both require that the biomass and protein copy number grows geometrically within individual cells. If this is true for proteins associated to replicating organelles, which accordingly display broader variances than other constitutive proteins (), there is no justification to assume that amounts of every constitutive protein grows geometrically in the cells. In addition to these a priori theoretical arguments the hypothesis that lognormal distributions are due to exponential growth can be more definitively dismissed based on the fact that identical distributions are observed in stationary populations of quiescent cells or in cell lines undergoing rapid proliferation (Figure1b), and that the number of divisions does not affect the breadth of the distribution (Figure1c).

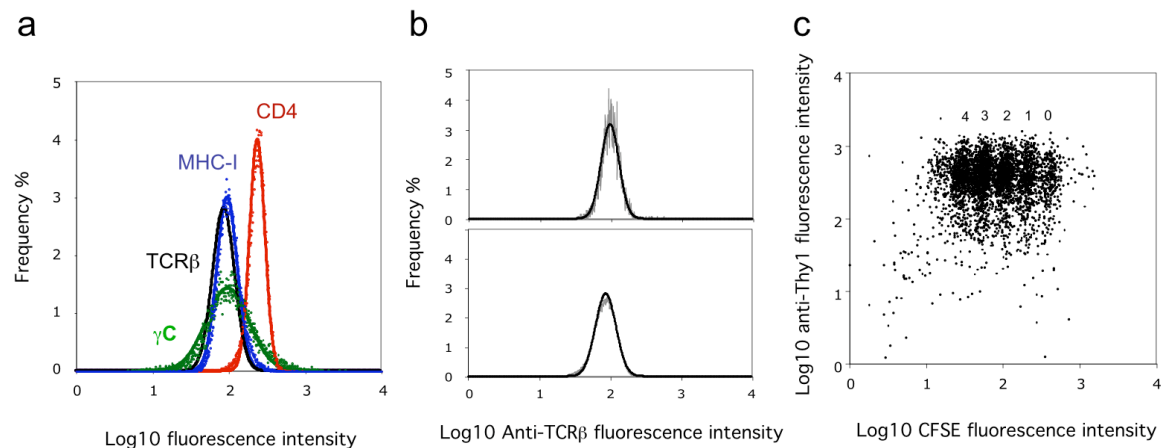


Figure 1: The copy number of several membrane proteins are distributed as a lognormal in both quiescent or exponentially growing T lymphocyte populations. A- Flowcytometry histograms of the expression of the indicated membrane proteins. B- The CD3 is distributed in the same way in quiescent T lymphocytes ex vivo and exponentially growing T cell lines. C- Serial division as assessed by CFSE delabeling does not change the protein variance in the population. Inset numbers are the number of divisions that the cells underwent

The above considerations indicate that lognormal distribution are not explained alone by stochastic effects inherent to the low numbers of molecules

neither by geometric propagation of small errors during cell division. The questions are: what is the nature of these extrinsic stochastic processes? Does the lognormal distribution contain any biologically meaningful information about the molecular machinery that produces it?

This article searches for sources of multiplicative noise within the molecular networks of quiescent cells that could explain why many constitutive proteins copy numbers are lognormal distributed in cell populations or tissues. We show that common motifs of biochemical networks are expected to produce lognormal distributions of protein copie numbers. This is due to the fact that regulation of the concentration of a given protein in the cell depends on multiple steps catalyzed by other cellular proteins or protein complexes (Rosen, 1991) introducing multiplicative noise. Furthermore, we show that the variance of specific protein content varies with the number of steps involved in its regulation, and that this variance might be a selectable quantitative trait, at least in multicellular organisms, where it could control the sensitivity of bulk cell population response curves.

Results

Many sources of multiplicative noise in metabolically stationary cells

In quiescent cells, as the vast majority of the cells in the body of multicellular organism, there is a continuous turnover of proteins. In these metabolically stationary cells the production and or degradation of a given protein species is often dependent on ongoing multistep biochemical processes. Some of these multistep processes are illustrated in Figure 2. It is clear that at least by statistical mechanics principles, they will be essentially stochastic in nature due to the finite number of molecules involved. The question is whether the organization of these processes introduces multiplicative noise that makes the final product display a log-normal distribution instead of distribution within Poisson family. The main result of this section is that for each of prototypic processes in the Figure 2 a rationale for the existence of multiplicative noise can be provided.

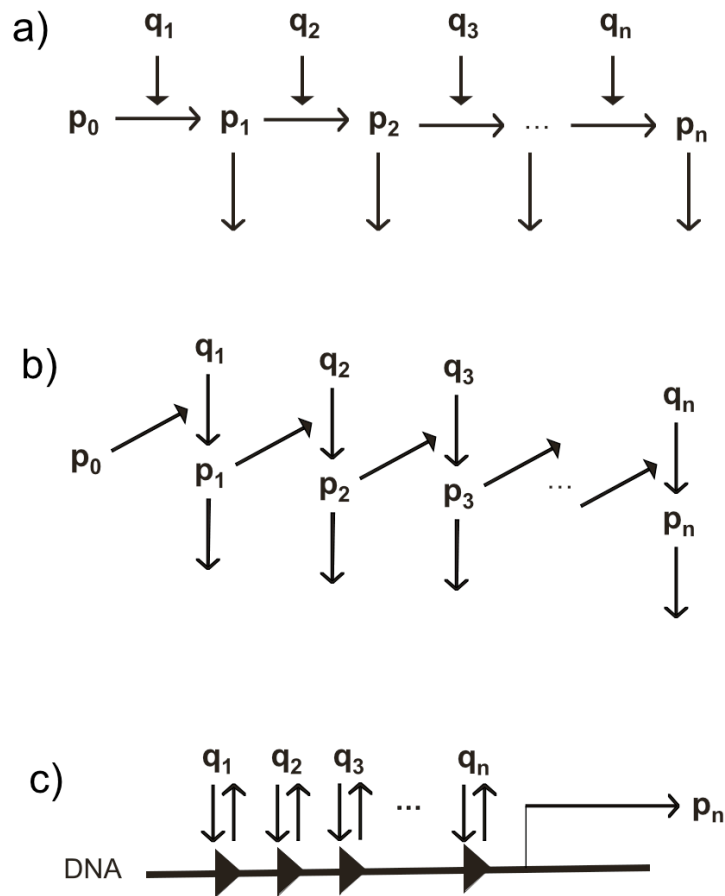


Figure 2: Three types of mechanism that exhibit multiplicative noise. a) a multistep biosynthetic chain, b) a regulatory cascade and c) the multiple transcription factor aggregation

The first prototype to be examined is the stepwise protein biosynthesis pathway. Following the emergence of newly synthesized polypeptidic chain the production and final localization of the mature protein follows many steps of modification of intermediate protein species. Amino- and carboxyl-terminal modifications, removal of signal sequences, modification of individual amino acids, addition of isoprenyl groups, addition of prosthetic groups, oligomerization with other proteins and formation of disulfide cross-links to name just a few posttranslational modifications. Quality control of all its subunits, mediated by interactions with chaperon proteins, also takes place providing extra steps the protein has to undergo. Moreover, the TCR is also trafficked across the cell, from the endoplasmatic reticulum to the membrane or organelles. All of these events contribute to the number of steps a protein has to undergo until it reaches the final functional stage. For example, the assembly of a fully functional T-cell receptor (TCR) (whose distribution is shown in Figure 1) needs the dimerization of two chains (alpha and beta), the subsequent association of this heterodimer with one CD3 gamma epsilon heterodimer, one CD3

delta epsilon heterodimer and an CD3 zeta zeta homodimer (Call et al., 2004). This complex needs to be trafficked to the intermediate cytoplasmic vesicle pool and family to the plasma membrane.

We model one such stepwise biosynthetic pathway assuming that the production rate of a given protein intermediate p_i is produced from the previous intermediate with a simple mass action kinetics under the action of a catalyst activity q_i , and that each intermediate can be lost or diverted to an alternative path with a first order kinetics with rate constant d_i . Under these conditions the derivative of the activity of each protein intermediate p_i is described by the following differential equation:

$$\frac{dp_i}{dt} = q_{i-1}p_{i-1} - d_i p_i - q_i p_i \quad (1)$$

If the catalyst activities q_i fluctuate slowly enough in time the activity of each protein i can be assumed to evolve in quasi-steady state. Under these conditions, the concentration of the final product of the chain p_n is given by :

$$p_n = p_0 \prod_{i=1}^n \frac{q_{i-1}}{d_i + q_i} \quad (2)$$

According to the central limit theorem $\log(p_n)$ the distribution of the final product in population tends to be normally distributed, provided that the erratic fluctuations of the catalysts of each step are uncorrelated.

The second prototype is a non-conservative cascade of biosynthetic or regulatory pathway, in which the intermediate protein or molecular species are different, synthesized de novo or modified from specific precursors. As compared to the above-mentioned biosynthetic pathway there is no conservation of the protein chain here. This kind of cascade is a feature of gene expression, transcription, and translation where the synthesis of many copies of RNA from one or two DNA segments, followed by the synthesis of many nascent protein chains from each mRNA, is concomitant with active “catalyzed” degradation of the mRNA and misfolded proteins by other molecular components of the cell. These cascades are also present in many signaling transduction cascades that control the level of specific proteins in the cells, for example the TCR illustrated above.

In multi-step non-conservative cascade the concentration of each element P_i can be described in its simplest form by the following differential equation:

$$\frac{dp_i}{dt} = q_i p_{i-1} - dp_i \quad (3)$$

which means that element p_i is produced with a rate proportional to the concentration of the previous element in the cascade p_{i-1} and to the activity of a catalyst q_i .

The quasi-steady state value of the final product n is:

$$p_n = p_0 \prod_i^n \frac{q_i}{d} \quad (4)$$

where p_0 is the first element in the cascade. Obviously if the activities of the precursors q_i fluctuate erratically in time and independently of each other, then we expect multiplicative noise to sneak in, such that the final product p_n will be log-normal distributed over the cell population.

The third prototype is the stepwise assembly of transcription factors necessary to recruit RNA polymerase, and therefore to transcribe a gene. If we consider that the rate of transcription depends on n transcription factors that all must be bound for transcription, we also get a process that propagates multiplicatively the noise in the activity of transcription factors. The dynamics of mRNA, m , is given by:

$$\frac{dm}{dt} = \prod_i^n b_i - \delta m \quad (5)$$

where δ is the decay rate of the RNA and b_i is probability that regulatory site i is occupied by the corresponding transcription factor q_i . Neglecting the noise introduced in all intermediate steps we write the equation for the corresponding protein p_n :

$$\frac{dp_n}{dt} = m - d_n p_n \quad (6)$$

where d_n is the decay rate of the protein. Each transcription factor can bind and be released from its regulatory site with rates q_i and d . If this process is in steady state the probability per time unit of finding the site i bound by its cognate factor is:

$$b_i = \frac{q_i}{q_i + d} \quad (7)$$

where q_i is the transcription factor binding activity assumed to be proportional to its concentration, and d is the dissociation rate. The concentration of the protein transcribed will be given by:

$$p_n = \frac{1}{\delta d_n} \prod_i^n \frac{q_i}{q_i + d} \quad (8)$$

This expression implies that there will be multiplicative noise propagation, provided that dissociation rate d is not negligible as compared to the binding activity q_i . Therefore, if the transcription factor binding activities q_i fluctuate in time or are different in each cell the protein p_n will be lognormal distributed in the cell population.

In summary, the three multistep biochemical processes modelled above are expected to display multiplicative noise propagation. To test whether this expectation

is fulfilled we simulated each model coupling the ordinary equations describing the n intermediate proteins p_i (according to eq. 1, eq. 3, or eq. 4-5, respectively) to stochastic differential equations describing the dynamics of each catalyst q_i . For simplicity, we assumed that each extrinsic catalyst q_i decays and is produced with constant rates. Under these conditions the catalyst activity is expected to undergo fluctuations in time displaying a Poisson distribution over time with mean $\langle q_i \rangle$ and standard deviation $\sqrt{\langle q_i \rangle}$. As a rule of thumb for more than 100 molecules the distribution is well approximated by a Gaussian distribution. Hence, we simulated the fluctuations of the catalysts using eqn. 9, a typical Langevin equation, that yields Gaussian fluctuations with average $\frac{k_q}{d_q}$ and standard deviation $\sqrt{\frac{k_q}{d_q}}$.

$$\frac{dq_i}{dt} = k_q - d_q q_i + \sqrt{\frac{k_q}{d_q}} \xi \quad (9)$$

ξ is Gaussian white noise with null mean and unitary standard deviation. In the three prototype models, and under some non-trivial parameter regimes, the concentration of the final product p_n was distributed over time and over the cell population as a log-normal as assessed by the Kolmogorov-Smirnov test, as can be seen on Figure 3. Furthermore, the standard deviation of the final product $\log(p_n)$ increases as a function of the number of steps in the process and also with the standard deviation of the extrinsic catalysts q_i , as can be seen on Figure 4.

Using our heuristic argument given above that the catalysts q_i fluctuate slowly enough for us to consider them constant and hence calculate the steady state of the system we can write the steady state of any intermediate step using a recursive relation of the general form:

$$p_{n+1} = F(q_n) p_n \quad (10)$$

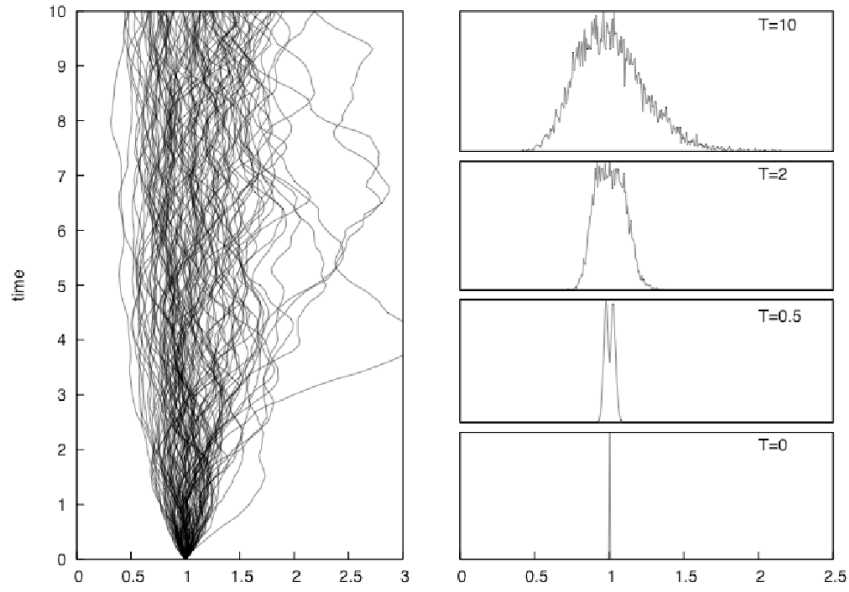


Figure 3- Fluctuation dynamics of the protein copies per cell and tendency towards an equilibrium standing distribution. The multistep protein biosynthesis chain model is illustrated using the following parameters: $n=10$; $k_q=0.1$, $d_q=0.001$, $d=10\langle q \rangle$.

The propagation of the variance along the chain depends critically on the characteristics of the function $F(p)$: as the slope of this function decreases the amount of variance that is propagated tends to zero. In order to guarantee that the final product follows the expected steady state the catalyst activities have to be smaller than the rate constants of the other processes. For example in conservative biosynthetic chain, the term $d+q_n$ has to be higher than q_{n-1} . Similarly, in the transcription complex assembly the dissociation rate d has to be higher than the binding activity q_i .

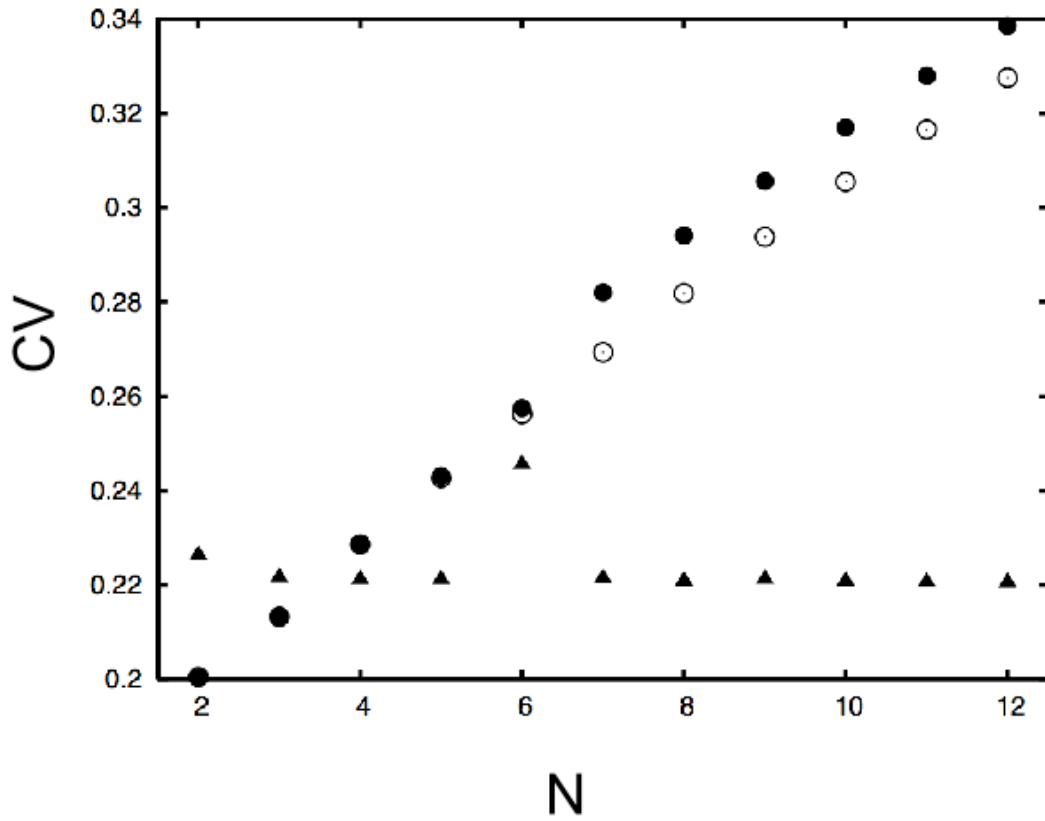


Figure 4- Dependence of the lognormal distribution on the structure of the underlying multistep process. Variation coefficient increases a function of the number of steps and of the noise introduced in each step. Parameters are the same as fig 3 except: (filled circles) the catalyst from step 6 to 7 has double the variance of the ones on fig 3 and triangles represent a chain with $d=0.1 \langle q \rangle$.

Log-normal distribution as an “attractor” of the stochastic biochemical networks in the cell

If catalyst activities undergo slow drifting fluctuations the final product of a biosynthetic process with sufficient steps is lognormal distributed. But can the actual noise structure of these fluctuations influence the shape of the final distribution? To address this question we studied the impact of the noise structure of the catalysts q_i by making them evolve according to the following Langevin equation:

$$\frac{dq_i}{dt} = f(q_i) + g(q_i)\xi \quad (11)$$

where the term ξ is a Gaussian white noise with mean zero, and f and g are functions chosen so that these q_i follow a predefined distribution. In order to achieve that, we

write the associated forward Kolmogorov equation (12) and solve it to the steady state obtaining equation 13. Then, by choosing the appropriate functions f and g , we produced a panel of input distributions to our protein assembly chains. The equations used and their respective asymptotic distributions are listed in table 1.

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial q_i} [f(q_i)\rho] + \frac{1}{2} \frac{\partial^2}{\partial q_i^2} [g^2(q_i)\rho] \quad (12)$$

$$\rho_{ss}(q_i) = \frac{A}{g^2(q_i)} \exp\left(-2 \int \frac{f(q_i)}{g^2(q_i)} dq_i\right) \quad (13)$$

where A is a normalization constant and ρ is the probability distribution function for the amount of catalyst.

Making the catalysts evolve accordingly to the equations on table 1, we integrated numerically the system of coupled equations in p_i and q_i as just described.

Table 1

Distribution	f	g^2
Gaussian	$k_q - d_q x$	σ
Gamma	$k_q - d_q x$	σx
Beta	$k_q - d_q x$	$\sigma x(1-x)$
LogNormal	$k_q - d_q \log(x)$	σx

The stationary distribution was recorded after transient effects have died out. This distribution was then compared to the lognormal using the Kolmogorov-Smirnov test. Irrespectively of the distribution used to generate the stochastic dynamics of the q_i 's, no significant difference to the lognormal was detected for all the prototypic multistep processes of length greater than a few steps, as can be seen on fig 3. These results were resilient to parameter changes. A beta distribution displaying a tail to lower values, scores in the first steps as the most distant from the lognormal distribution but still converges back to the latter as the number of steps increases (Fig 4).

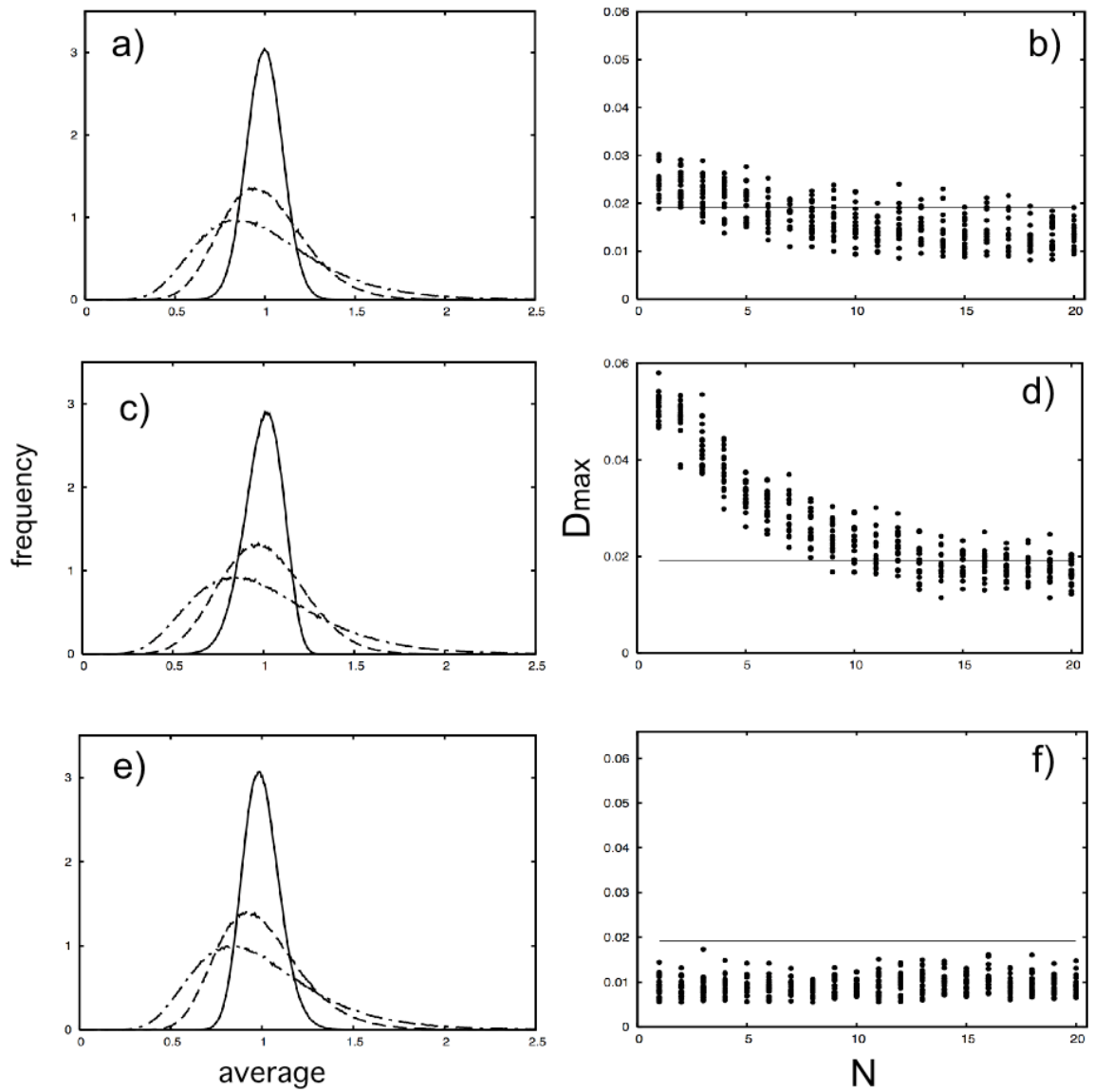


Figure 5: The convergence to the lognormal does not depend on the catalyst fluctuation noise structure. Gaussian (a,b), beta (c,d) and gamma (e,f) distributed fluctuations of catalysts as a function of the number of steps for the multistep biosynthetic chain. Left panel display the frequency distribution for the indicated number of steps. Right panel: Maximal distance between the cumulative density function and a lognormal with the same mean and variance for many simulations. Below the line the distribution is statistically indistinguishable from the lognormal according to the Kolmogorov-Smirnov test. Parameters are the same as Figure 3 except the ones controlling the q_i dynamics which were chosen such that these distributions would have the same characteristics as the gaussian: $kq=10$ $dq=14$ $\sigma=2$ (beta), $kq=0.92$ $dq=0.2$; $\sigma=0.0001$ (lognormal)

Variance of protein copy number per cell as a selectable quantitative trait

In contrast with the models of (Furusawa et al., 2005) and (Krishna et al., 2005), where the standard deviation of the protein copy number is predicted to be proportional to its average value for any cellular protein, the three models proposed here uncouple these two properties of the distribution. Thus, the standard deviation scales with the number of steps in the process and with the amplitude of noise fluctuations of each catalyst, independently of the mean value of the final product. Therefore, according to our model the variance of the lognormal is controlled by the structure and dynamics of the multistep process and thus could be a quantitative trait upon which natural selection could potentially operate. The question is whether this trait is of any functional relevance? The sensitivity of the response curves of biochemical, signaling, and genetic pathways has been closely scrutinized by both theoretical (Heinrich et al., 2002) and experimental (Raser and O'Shea, 2004) studies, and shown to be of evolutionary relevance. As we will now demonstrate, the variance of specific protein copy number per cell has a direct bearing on the sensitivity of the bulk response of a cell population, and therefore could be a meaningful quantitative phenotype of a cell population or tissue.

It is well known within biochemistry, cell biology, or cellular immunology practices that graded bulk responses of a cell population to a stimulus might be obtained if all the cells make a graded response or alternatively if only a fraction of the cells engage in a yes-or-no response (Ferrell and Machleder, 1998). In this latter scenario, graded responses to increasing stimuli are obtained by progressive increase in the fraction of responding cells, as more and more cells receive a suprathreshold stimulus. Obvious examples are graded proliferative or apoptotic responses involving molecular switches or check points. But similar binary responses have been documented for the other functions, namely cytokine production (Ko et al., 1990; Hollander et al., 1998; Bayley et al., 2003; Bayley et al., 2003; Guo et al., 2005; Calado et al., 2006). Interestingly, there is a direct quantitative correspondence between the mean and variance of the log-normal distribution of a protein whose activity defines a response threshold and the parameters of a Hill-type response curve (eqn 14) typically used to assess cooperative/sensitivity.

$$\text{Normalized bulk response} = R(s) = \frac{s^h}{k^h + s^h} \quad (14)$$

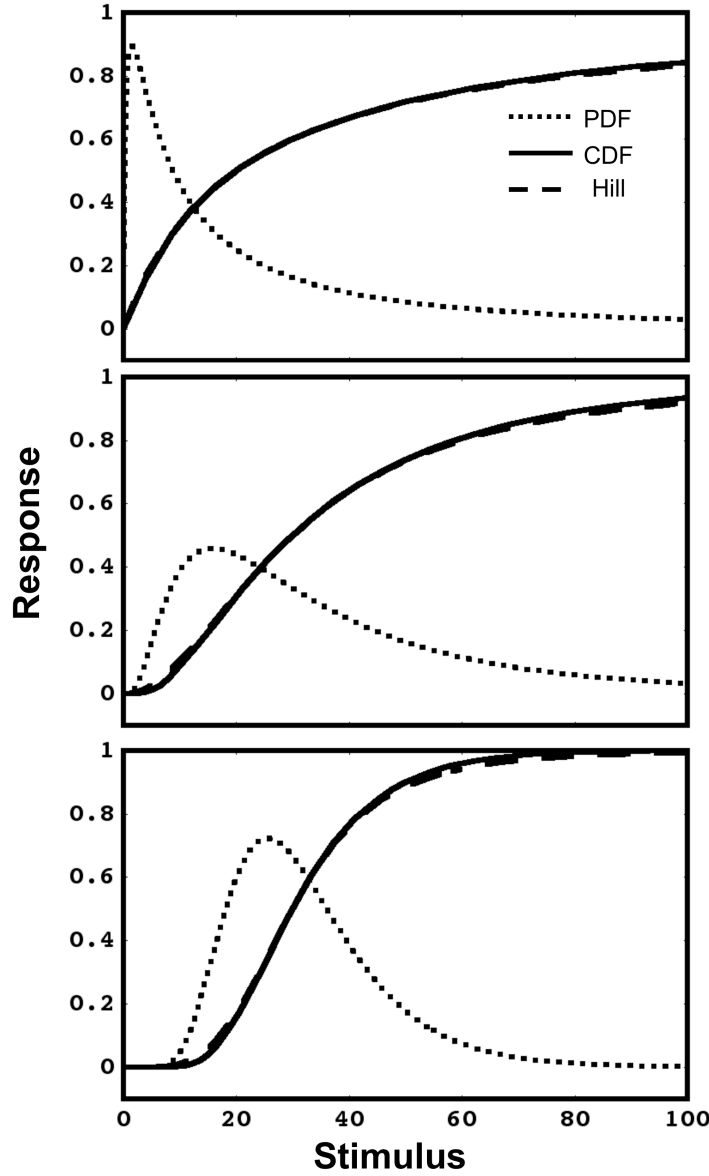


Figure 6: Cumulative response curves of a population whose threshold is distributed as a lognormal (fine dash) and its comparison with a Hill function (continuous line). A Hill function is mapped into a cumulative lognormal distribution (long dashes) with $M=\log(k)$ and

$$S = \frac{2}{h} \sqrt{\frac{2}{\pi}}$$

. a) Hill function with $h=1$, $k=20$. b) $h=2$, $k=30$ c) $h=4$, $k=30$.

Consider a scenario in which the threshold for a binary cellular response to a graded stimulus is set by a protein whose copy number p is described by any of the multistep processes modeled above. In each individual cell the p does an independent slow random walk with the properties previously described, and accordingly a “snapshot” of the cell population at any given time point would display lognormal distribution of p . Since the bulk response of the population is the sum of

the responses of the individual cells (assuming for simplicity that there are no synergies stemming from intercellular communication) we can write:

$$R(s) = N \int_0^{\infty} H(s - p) \rho(p) dp \quad (15)$$

Where $\rho(p)$ is the probability density function of the threshold protein copy number p ; $r(s, p)$ is the binary response of the individual cell assumed to be a Heaviside function that takes unitary value for a supra-threshold stimulus ($s > p$), and is zero valued otherwise; and N is a normalization constant. Under these conditions, the expression (eqn 15) can be rewritten as the cumulative density function (CDF) of threshold protein p in the cell population.

$$R(s) = \int_0^s \rho(p) dp \quad (16)$$

Imposing that the median of lognormal distribution of the threshold protein $\rho(p)$ is identical to the half-maximal stimulus of the Hill function k , and that the derivative of the two functions $R(s)$ is the same at $k=s$ this point we reach the following expressions that map one function to the other.

$$\begin{aligned} k &= \exp(M) \\ h &= \frac{\sqrt{8}}{\sqrt{\pi} S} \end{aligned} \quad (17)$$

The two functions are plotted for several parameter values in Figure 6. These expressions (eq. 17) are a biunivocal mapping of the lognormal distribution to the any Hill-type response curve. Thus, by changing the breadth of the underlying lognormal one can recover simple Michaelis-Menten curves as well as sigmoidal curves with apparent positive and negative cooperativity. We interpret this result as a strong suggestion that, at least in multicellular organisms, the variance of the lognormal distribution of the protein copy number in a cell population or tissue is a relevant phenotype, since it controls that sensitivity of a tissue response.

Discussion

The copy numbers of many constitutive proteins are lognormal distributed in isogenic prokaryotic or eukaryotic cells populations, irrespective of whether these are quiescent or undergoing exponential growth. These lognormal distributions can be generated by multistep protein biosynthetic processes, which, under nontrivial conditions, propagate intrinsic noise in a multiplicative way. Since these biosynthetic processes are pervasive in the cellular genetic and metabolic networks the lognormal

distribution is expected to be an attractor for the constitutive stochastic dynamics of the cell, and thus to be ubiquitously found.

Multiplicative noise might seem naively a dangerous property that should have been eliminated from the organization of the cellular machinery by natural selection. Since in a biochemical or genetic networks every component is connected to all other components to higher or lesser extent (Barabasi and Oltvai, 2004) one can find by recursion chains of arbitrary length. If noise would be always propagated multiplicatively the whole variance would either explode or collapse. A solution to this puzzle is, however, suggested by the simulations of the three models for stochastic protein expression. The results indicate that there are parameters values where the variance in each step contributes to the final step, while there are other parameters where the variance is dampened (Figure 5). For example, the reduction of the decay and divergence rate in a biosynthetic modification pathway eliminates the noise introduced by the each step due to mass conservation that naturally compensates perturbations instead of propagating them. Similarly, in the model of stepwise transcription factor assembly there are binding activities and dissociation rates of each factor that propagate noise while others do not, the two extremes being all the binding sites being either free most of the time or bound most of the time. This means that the global topology and operation concentrations of the multistep processes can control the propagation of the variance through the network, avoiding unbounded recursive noise amplification, even without need to call into action additional feedbacks.

Multiplicative noise propagation is not trivially obtained in the simple multistep biochemical models. In particular, propagation of noise in the protein synthesis chain model requires that there is protein loss in each step, otherwise the noise is canceled. The magnitude of these loss facilitates the convergence of the chain to the lognormal distribution, and the larger it is the faster the convergence. This raises an issue on the efficiency and yield of protein expression pathway in which the proteins are lognormal distributions. More than 50% of the nascent proteins are misfolded and targeted for destruction (Turner and Varshavsky, 2000), and thus significant waste is not unlikely in a protein expression pathway particularly when, as is the case of the TCR, it involves several multimerization steps and trafficking events. The other prototype models propagate noise under parameters that are less surprising. The multistep cascade is very robust to parameter changes, while the multiple transcription factor aggregation model requires a handful of independent and reversible binding events, which are very plausible constraints.

Alternative explanations for the ubiquity of the lognormal distribution of specific protein copy number per cell (Furusawa et al., 2005; Krishna et al., 2005), which rely on autocatalytic growth of the cell mass, predicted a general linear relationship between the mean and the variance of the distribution. High throughput single cell proteomic analysis of hundreds or thousands proteins indicates that such trend can be observed for inducible gene products during exponentially growing cell populations submitted to a challenge (Bar-Even et al., 2006). However, a similar analysis of constitutive protein distributions in cell lines derived from higher organisms indicated that each protein has a characteristic copy number variance across the population (Sigal et al., 2006). Together with the observation that the same standing copy number variation is observed in both expanding cell lines and quiescent cells (Figure 1), these observations indicate that the mechanism generating the lognormal distribution can uncouple its average and variance, as the models proposed here in which the two moments can be set in a virtually independent way. Our results are therefore in line with these findings.

It has been appreciated before that bulk measurement in cell populations could mask switch-type binary responses at the single cell level. One of the characteristics of the lognormal distribution that underlies the variation in many constitutive membrane receptors across cell populations is that any bulk response described by a Hill-type function can be interpreted as the cumulative response of individual cells with a lognormal distributed response threshold (Figure 6). In biochemical, signal transduction, or gene regulation modeling it is common to fit Hill-type functions to cell response curves. From the Hill exponent one infers negative or positive cooperativity, following the well established examples from adsorption, and tries to infer the underlying biochemical network features that could explain this degree of cooperativity. The biunivocal mapping of Hill response to a binary response dependent on lognormal distributed thresholds implies that caution in the interpretation of the bulk response curves, and that inferring positive or negative cooperative processes might be unwarranted. This is particularly significant with the widespread use of large scale transcriptomic and proteomic tools that measure the changes in time of thousands of gene products but that require measuring the sum of the contribution of many cells. Inferring the topology of the regulatory networks from these average kinetics might be misleading (as discussed in chapter IV of this thesis).

Control of the width of the response threshold distribution over a cell population or tissue might be the means whereby a particular collective response curve to a stimulus is achieved in a multicellular organism. The distribution of response thresholds increases the sensitivity of the whole cell population to a wider range of

stimuli and in the case of the response of a precursor population to differentiation signals, a wider distribution of the threshold might facilitate the maintenance of precursor pool, which eventually could be fully committed if the distribution would be too sharp. The variance of the distribution could be therefore tuned to perform specific function by natural selection of variants with different number of steps (Alves et al., 2002) or the relationships between kinetic parameters (Kunin et al., 2004). If collective cell responses are a legitimate phenotype of multicellular organisms, the same does not apply to unicellular organisms such as bacteria and yeast, where lognormal distribution of proteins may be a functionless consequence of molecular machinery organization.

Methods

Stochastic differential equations of the form $\frac{dy(t)}{dt} = f(y(t)) + g(y(t))\xi$, where ξ is

Gaussian white noise, were solved with the stochastic Euler method:

$$y(t + \Delta t) = y(t) + \Delta t f(y(t)) + \sqrt{\Delta t} g(y(t)) \xi,$$

with a step size $\Delta t = 10^{-4}$.

Comparison with the lognormal distribution was made collecting long time series generated by the SDE's. The time series was subsampled to 8000 points in order to remove any autocorrelation. These data were then tested for their goodness-of-fit to a lognormal distribution with the same average and standard deviation using the Kolmogorov-Smirnov test with a significance of 5%.

References

- Alves, R., R. A. Chaleil and M. J. Sternberg (2002). "Evolution of enzymes in metabolism: a network perspective." *J Mol Biol* **320**(4): 751-70.
- Bar-Even, A., J. Paulsson, et al. (2006). "Noise in protein expression scales with natural protein abundance." *Nat Genet* **38**(6): 636-43.
- Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." *Nat Rev Genet* **5**(2): 101-13.

- Bayley, J. P., A. M. Bakker, et al. (2003). "Analysis of allelic expression patterns of IL-2, IL-3, IL-4, and IL-13 in human CD4+ T cell clones." Eur J Immunol **33**(8): 2142-8.
- Bayley, J. P., J. G. van Rietschoten, et al. (2003). "Allele-specific expression of the IL-1 alpha gene in human CD4+ T cell clones." J Immunol **171**(5): 2349-53.
- Blake, W. J., K. A. M., C. R. Cantor and J. J. Collins (2003). "Noise in eukaryotic gene expression." Nature **422**(6932): 633-7.
- Calado, D. P., T. Paixao, D. Holmberg and M. Haury (2006). "Stochastic monoallelic expression of IL-10 in T cells." J Immunol **177**(8): 5358-64.
- Call, M. E., J. Pyrdol and K. W. Wucherpennig (2004). "Stoichiometry of the T-cell receptor-CD3 complex and key intermediates assembled in the endoplasmic reticulum." Embo J **23**(12): 2348-57.
- Carneiro, J., T. Paixão, et al. (2005). "Immunological Self-Tolerance: Lessons from Mathematical Modeling." J Comp Appl Math **184**: 77-100.
- Chen, B. S. and Y. C. Wang (2006). "On the attenuation and amplification of molecular noise in genetic regulatory networks." BMC Bioinformatics **7**: 52.
- Doncic, A., E. Ben-Jacob and N. Barkai (2006). "Noise resistance in the spindle assembly checkpoint." Mol Syst Biol **2**: 2006 0027.
- El-Samad, H. and M. Khammash (2006). "Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks." Biophys J **90**(10): 3749-61.
- Ferrell, J. E., Jr. and E. M. Machleder (1998). "The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes." Science **280**(5365): 895-8.
- Furusawa, C., T. Suzuki, et al. (2005). "Ubiquity of log-normal distributions in intracellular reaction dynamics." Biophysics **1**: 25-31.
- Guo, L., J. Hu-Li and W. E. Paul (2005). "Probabilistic regulation in TH2 cells accounts for monoallelic expression of IL-4 and IL-13." Immunity **23**(1): 89-99.
- Heinrich, R., B. G. Neel and T. A. Rapoport (2002). "Mathematical models of protein kinase signal transduction." Mol Cell **9**(5): 957-70.
- Hollander, G. A., S. Zuklys, et al. (1998). "Monoallelic expression of the interleukin-2 locus." Science **279**(5359): 2118-21.
- Kaern, M., T. C. Elston, W. J. Blake and J. J. Collins (2005). "Stochasticity in gene expression: from theories to phenotypes." Nat Rev Genet **6**(6): 451-64.
- Ko, M. S., H. Nakauchi and N. Takahashi (1990). "The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates." Embo J **9**(9): 2835-42.

- Krishna, S., B. Banerjee, T. V. Ramakrishnan and G. V. Shivashankar (2005). "Stochastic simulations of the origins and implications of long-tailed distributions in gene expression." Proc Natl Acad Sci U S A **102**(13): 4771-6.
- Kunin, V., J. B. Pereira-Leal and C. A. Ouzounis (2004). "Functional evolution of the yeast protein interaction network." Mol Biol Evol **21**(7): 1171-6.
- McAdams, H. H. and A. Arkin (1997). "Stochastic mechanisms in gene expression." Proc Natl Acad Sci U S A **94**(3): 814-9.
- McAdams, H. H. and A. Arkin (1999). "It's a noisy business! Genetic regulation at the nanomolar scale." Trends Genet **15**(2): 65-9.
- Milutinovic, D., J. Carneiro, M. Athans and P. L. Lima (2007). "Modeling Dynamics of Cell Population Molecule Expression Distribution." NonLinear Analysis Hybrid System **1**: 81-94.
- Ozbudak, E. M., M. Thattai, et al. (2002). "Regulation of noise in the expression of a single gene." Nat Genet **31**(1): 69-73.
- Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-4.
- Rosen, R. (1991). Life Itself, Columbia University Press.
- Sigal, A., R. Milo, et al. (2006). "Variability and memory of protein levels in human cells." Nature **444**(7119): 643-6.
- Thattai, M. and A. van Oudenaarden (2001). "Intrinsic noise in gene regulatory networks." Proc Natl Acad Sci U S A **98**(15): 8614-9.
- Turner, G. C. and A. Varshavsky (2000). "Detecting and measuring cotranslational protein degradation in vivo." Science **289**(5487): 2117-20.



Stochastic Monoallelic Expression

Quantitative insights into stochastic monoallelic expression of cytokine genes

Tiago Paixão¹, Tiago P. Carvalho¹, Dinis P. Calado¹, and Jorge Carneiro^{1,2}

1. Instituto Gulbenkian de Ciência, Oeiras, Portugal

2. Associate Laboratory of Oeiras, Portugal

Summary

Gene expression from both parental alleles is beneficial by masking the effects of deleterious recessive mutations, and reducing the noise in gene expression in diploid organisms. However, a class of genes are expressed preferentially or strictly from a single allele. The selective advantage of avoiding biallelic expression is clear for allelically excluded antigen receptor and odorant receptor genes, genes undergoing X-chromosome inactivation in females, and parental genomic imprinted genes. In contrast, there is no clear biological rationale for the predominant and stochastic monoallelic expression of cytokine genes in the immune system, and the underlying mechanism is elusive and controversial. A clarification of the mechanism of predominant monoallelic expression would be instrumental in better understanding its eventual biological function. This prompted the development of a quantitative framework that could describe the dynamics of the pattern of allele expression of the IL-10 gene, from which general quantitative insights could be gained. We report that the experimental observations on these patterns of allelic expression cannot be easily reconciled with a simple model of stochastic transcriptional activation, in which the two alleles are, at any time, equally competent for transcription. Instead, these observations call into action a general model of eukaryotic transcriptional regulation according to which the locus competence for transcription is dynamic, involving multiple, cooperative and stochastic modification steps. In this model, the probability that an allele becomes transcriptionally active is a function of the number of chromatin modifications that it accumulated. Based on the properties of this model, we argue that predominant monoallelic expression might have had no adaptive role, and may have evolved under indirect selection for low frequency of expressing cells.

Introduction

Gene expression from both parental alleles is beneficial by masking or minimizing the effects of deleterious recessive mutations in diploid organisms (Otto and Goldstein, 1992). Biallelic expression also contributes for the robustness of diploid cells by reducing the intrinsic noise in gene expression (Cook et al., 1998). Surprisingly, a growing class of genes is expressed preferentially from a single allele in each cell (Goldmit and Bergman, 2004). In the most well studied cases, the functional advantage of monoallelic expression is clear. Antigen receptor genes (Held et al., 1995; Spicuglia et al., 2006) and odorant receptor genes (Serizawa et al., 2004) are isotypic and allelic excluded, resulting in a “one cell-one receptor rule” that is essential for recognition specificity in the immune and nervous systems, respectively. One of the two X-chromosomes in every mammalian female cell is silenced by a long range, monoallelic gene silencing process that mediates essential gene dosage compensation between the two sexes (Heard and Disteché, 2006). Some autosomal genes, the so-called imprinted genes, are expressed in a parent-of-origin fashion that plays a key role in regulating offspring development and mother-offspring interactions (Wolf and Hager, 2006), and in mediating sex genetic conflicts (Day and Bonduriansky, 2004). In contrast with the above examples, the functional benefit of monoallelic expression of some autosomal genes, such toll-like receptor 4 gene (Pereira et al., 2003), and several cytokines genes (Bix and Locksley, 1998; Hollander et al., 1998; Kelly and Locksley, 2000; Calado et al., 2006) is less evident.

The pattern of allelic expression of cytokines genes by cells of the immune system is particularly intriguing. The genes encoding IL-2 (Hollander et al., 1998), IL-4 (Bix and Locksley, 1998; Riviere et al., 1998), IL-5 (Kelly and Locksley, 2000), IL-13 (Kelly and Locksley, 2000) and IL-10 (Calado et al., 2006) are predominantly monoallelic expressed, but not “allelically excluded”. Thus, in T cell populations that have been cultured *in vitro* under conditions that drive upregulation of these cytokine gene, and even in cell lines or clones selected for specific cytokine secretion, only a fraction of the cells expresses the cytokine gene at any given moment, and among these cells a small proportion does it biallelically. The proportions of cells that do not express the gene, that express it from a single allele, and from both alleles are compatible with a stochastic mechanism where the two alleles are expressed independently with a probability that is less than one. Moreover, cells that did

express and cells that did not express the IL-4 gene, when subsequently cultured, had the same probability of expressing the gene (Hu-Li et al., 2001; Guo et al., 2004). The same was confirmed for the IL-10 gene (Calado et al., 2006), indicating that the patterns of allelic expression of these two cytokines are unstable at single cell level, but remarkably robust in cell populations or clones (Hu-Li et al., 2001).

What is then the mechanism underlying such stochastic predominant monoallelic expression of cytokines? What determines the probability that an allele becomes expressed? The simplest, almost trivial, explanation is that the two alleles undergo random uncorrelated transcriptional bursts, which only seldomly coincide (resulting in biallelic expression), due to low number of transcription factors (Kuznetsov et al., 2002). A more elaborate hypothesis, is that the two allelic loci are differently accessible to the transcription factors, because they undergo independent epigenetic chromatin modifications in each cell. The available experimental evidence does not allow discriminating between these two hypotheses. Asynchronous DNA replication reflects the differential epigenetic modifications of the alleles of a gene and was observed in all instances of monoallelic expression (Gimelbrant and Chess, 2006). Accordingly, the IL-4 gene was shown to replicate asynchronously (Singh et al., 2003), implicating different epigenetic states of the alleles. Furthermore, linked monoallelically expressed genes also undergo coordinated asynchronous replication (Singh et al., 2003). IL-4, IL-13 and IL-5 genes, which are linked and in close proximity in the chromosome, were reported to be coordinately expressed from either of the chromosomes (Kelly and Locksley, 2000), suggesting the action of some broad range chromatin remodeling mechanism. However, a recent reexamination of coexpression of IL-4 and IL-13 genes (Guo et al., 2005) revealed that the four alleles were independently expressed. Although this observation does not rule out the possibility that chromatin remodeling operates independently at the two closely located loci, it seems to favor the more parsimonious hypothesis of uncorrelated transcriptional bursts.

Aiming to gain quantitative insights into the mechanism of predominant monoallelic expression of cytokine genes we modeled the two candidate mechanism above. We report here that the experimental observations on the patterns of allelic expression of the IL-10 gene cannot be easily reconciled with a simple model of stochastic transcriptional activation, in which the two alleles are, at any time, equally competent for transcription. Instead, these observations call into action a mechanism in which locus competence for transcription is dynamic, involving multiple, cooperative and stochastic modification steps. In this model, the probability that an allele becomes transcriptionally active is a function of the number of chromatin

modifications that it accumulated. Based on the properties of this model, we argue that predominant monoallelic expression might have no adaptive significance, and has evolved under indirect selection for low frequency of expressing cells.

Results

Stochastic monoallelic IL-10 gene expression: allele independence, low expression probability and instability at the single cell level.

In this article we will derive two models and assess their capacity to account for the observations on allelic expression pattern of IL-10 locus, previously described by Calado et al. (Calado et al., 2006) (Figure1a). An IL-10 reporter transgenic mouse was generated by straightforward knock in approach, in which the coding region was substituted by that of yellow fluorescent protein (YFP), thus generating a new allele, denoted IL-10^{KI}. The product of this allele can be easily measured and distinguished from that of the wildtype allele (IL-10^{WT}) in cells from heterozygous animals. Splenic CD4+ T cells from heterozygous IL-10^{WT/KI} mice become IL-10 expressors after being stimulated with plate-bound anti-CD3 and anti-CD28 in the presence of dexamethasone and vitamin D3. Monoallelic expression of the IL-10 gene products predominates over biallelic expression (Figure1a). Analysis of the frequency of cells expressing either IL-10^{WT} and IL-10^{KI} alleles alone, or both alleles indicates that the probability of expressing any of the allele is independent (Fisher exact test) and about 10%. This probability can be increased by increasing the magnitude of the TCR stimulus, but remains always stochastic and independent. More important, for the purpose of the present analysis, the pattern of allele expression is only stable at the population level. Thus, when highly purified IL-10^{KI} positive cells were stimulated for 6 additional days, under the same conditions complemented with IL-2, the original pattern of expression was slowly restored (Figure1b). In the next sections we will ask the models to reproduce the probability $p=10\%$ that alleles are expressed as well as the relaxation time $t_{relax}>5$ days of the purified populations. Notice that by assuming an identical probability for both alleles we are neglecting the consistent observations that the probability of expression or detection of the product of the artificial allele is slightly lower than that of the natural IL-10 allele.

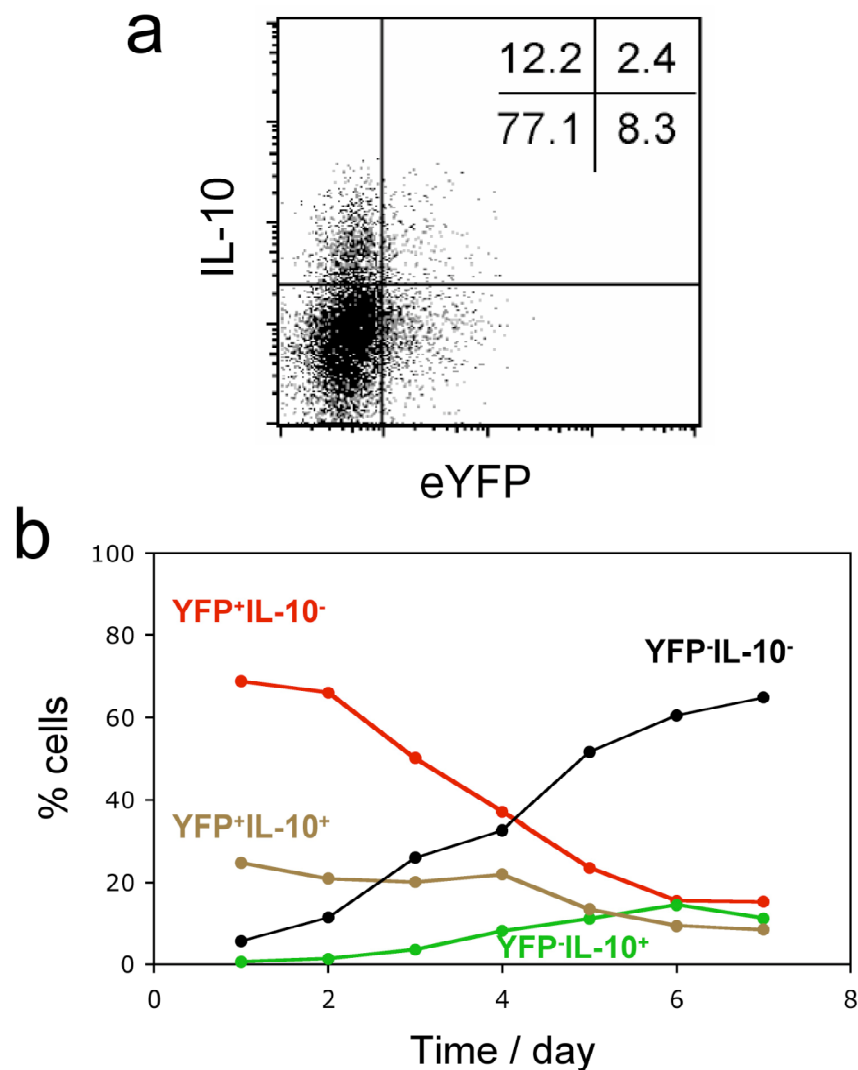


Figure 1- Stochastic predominant monoallelic expression of IL-10 gene products in CD4⁺ T cells. A) Splenic CD4⁺ T cells from heterozygous transgenic IL-10^{WT}/IL-10^{KI} mice were MACS purified and stimulated with plate-bound anti-CD3, and anti-CD28 in the presence of Dexamethasone and Vitamin D3. At day 6, cells were stained intracellularly for IL-10^{KI} (YFP) and IL-10^{WT} gene products and analyzed by flow cytometry. Numbers represent relative %. Data representative of at least 2 independent experiments. B) CD4⁺ T cells expressing the artificial YFP allele (constituted by YFP+IL-10⁻ and YFP+IL-10⁺ cells) were sorted to 99% purity from the previous 6 day stimulated CD4⁺ T cells culture, and re-stimulated as before in the presence of mouse IL-2. The graph represents the time course of the percentages of the indicated subpopulations during the restimulation protocol.

Simple stochastic transcription activation predicts predominant monoallelic expression but under unrealistic parameter regimes

We adopted the simple model by Cook et al. (Cook et al., 1998) describing the stochastic allelic expression according to the reaction diagram of Figure 2. The key feature of this model is that the transcription activation rate is the same for the two alleles at any time point, as it is assumed in most models of gene expression. Each allele switches independently and stochastically between transcriptionally inactive and active states as the transcription factors necessary for attracting the RNA polymerase associate and dissociate at its promoter. Denoting transcriptional activation and deactivation rates by a and d , respectively, the time intervals each allele spends in transcriptionally inactive and active states are exponentially distributed with mean $1/a$ and $1/d$, respectively.

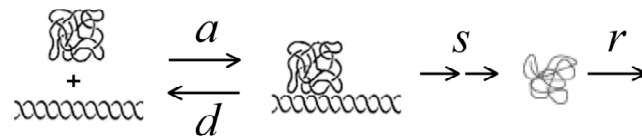


Figure 2- Simple stochastic model of gene expression based on Cook et al. (Cook et al., 1998). The model follows for individual alleles of a gene and the concentration of its product x . Each allele switches between transcriptionally inactive and active states with rates a and d respectively. While in the active state each allele transcribes and translates its product whose concentration increases with rate s , and decreases with rate r . (this illustration is inspired by the cartoons by J. Monod).

When the allele is transcriptionally active the corresponding protein is produced with a net rate s , aggregating transcription and translation. The newly produced protein is released to the medium by exocytosis and/or is degraded with rate r . Allele x product concentration is therefore described by the following piecewise equation:

$$\begin{cases} \frac{dx}{dt} = s - rx & \text{if the allele is transcriptionally active} \\ \frac{dx}{dt} = -rx & \text{otherwise} \end{cases} \quad (1)$$

The time evolution of the allele expression pattern of a single cell is obtained by running two independent Monte-Carlo type simulations of the stochastic serial switching between active and inactive states and integrating eqn 1 in piecewise manner (illustrated in Figure 3).

In the experiments, cells were categorized as positive or negative for each allele (Figure 1), and therefore we need to define a relationship between the model's parameters and variables and the frequencies of allele positive cells. In a first approach we neglected the protein concentration altogether and assumed that the cell is positive for an allele that is transcriptionally active, and is otherwise negative for that allele. The probability that the allele is active at any time point is:

$$p_{act} = \frac{a}{a + d} \quad (2)$$

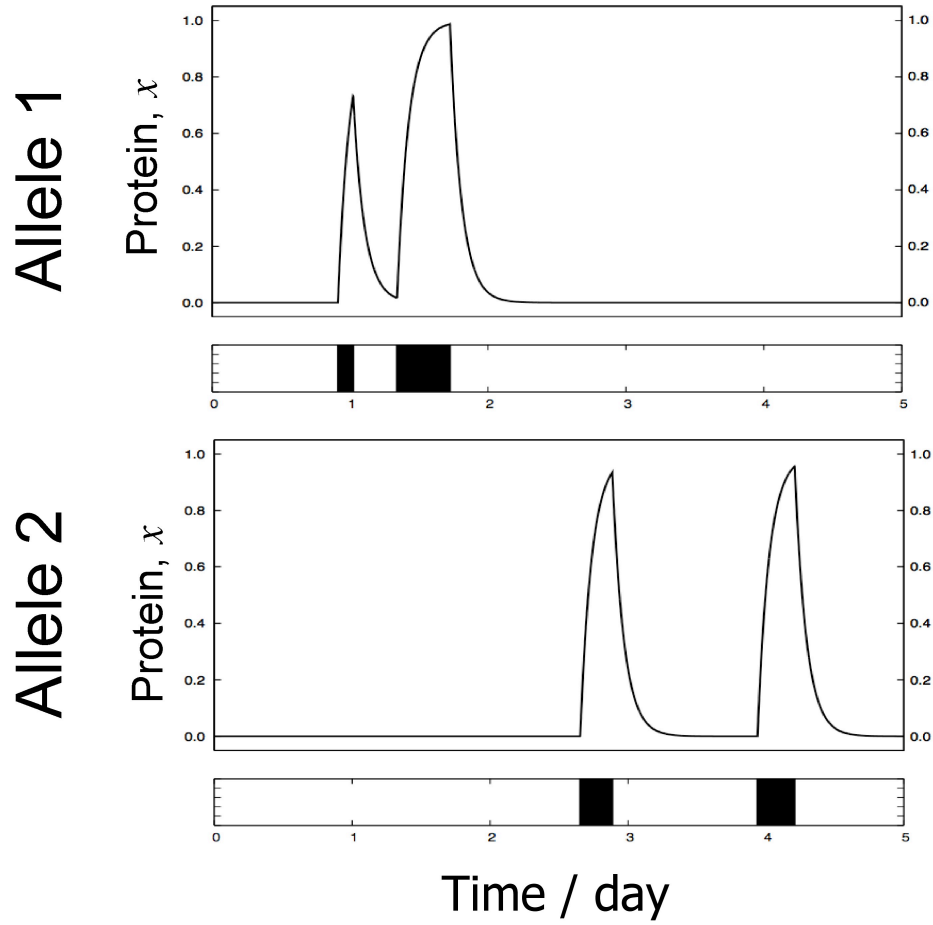


Figure 3- Illustration of a time series produced by the minimal mechanism of stochastic transcription. For each allele, the top graph represents the evolution of the protein concentration normalized so that the steady state predicted for permanent transcriptional activation is unitary (line) and the bottom graph represents the transcriptionally active (black) and inactive (white) states. Parameter values: $a= 0.00044 \text{ min}^{-1}$; $d=0.004 \text{ min}^{-1}$; $s=0.0083 \text{ au min}^{-1}$; $r=0.0083 \text{ min}^{-1}$.

As stated in the previous section the value of this probability is about 10% for IL-10 locus, which requires setting $d=9a$. Provided that this condition is fulfilled, the predicted pattern of allele expression at the population level will be in good agreement with the observed frequencies. The switching rate between transcriptionally active and inactive states will be compatible with observed relaxation time of the population, provided the average time the alleles stay inactive (that precedes the slower transition) is of same magnitude, i.e. $1/a \approx t_{relax} > 5$ days (notice that the simulation illustrated in Figure3 is faster than this time scale, randomizing with $1/a \approx 2$ days). This value seems too large. In fact, typical association times of transcription factors and DNA *in vitro* (often measured at 25°C) range from 2 to 150 minutes (Affolter et al., 1990; Ekker et al., 1991; Trieu et al., 1999; Kwon et al.,

2001). Since transcription factors are necessary to sustain the transcriptionally active state the association times provide rough estimates of the order of magnitude of the realistic values of $1/d$. This would imply that the inactive transcription periods in the model would be 9x longer, i.e. between half hour and one day. The turnover of transcription factors associated to target sites in chromosome as measured by FRAP in living cells provides another, perhaps more accurate estimate of the association kinetics. The kinetics of association and dissociation measured in this way are even faster than the previous ones, occurring in the time scale of seconds to minutes in transcriptionally active chromatin (Yao et al., 2006), and even in mitotic chromatin (Chen et al., 2005). This seems to warrant the conclusion that the model estimates of transcriptionally activate and inactive periods of the IL-10 gene are unrealistically long.

Would a consideration of the dynamics of allele product concentration change this conclusion? From equation 1 and through the same analysis described in (Carneiro et al., 2005) we can derive the probability density function (PDF) for the allele product concentration x in a population of cells, denoted $\rho(x_i)$. This is given by the following expression:

$$\rho(x) = x_i^{\frac{a}{r}-1} (s - r \cdot x_i)^{\frac{d}{r}} \quad (3)$$

This PDF corresponds to a scaled beta distribution with canonical parameters $\alpha=a/r$ and $\beta=d/r+1$. Since $\beta>1$ for all parameter values, this distribution can only assume two different shapes depending on the value of α : a strictly decreasing shape, if $\alpha<1$, and an unimodal shape, if $\alpha>1$. These two different parameter regimes are related to the relative timescales between the protein degradation and the allele activation rate. If $a<r$ ($\alpha<1$) the protein dynamics follows the allele activation dynamics and closely reflects the state of the allele. On the other hand, if $a>r$ the protein dynamics are slower than the allele dynamics and the protein can persist between transcription bursts. As the most likely value of the allele product concentration in this distribution becomes positive, the probability of having the biallelic expression at the protein level increases. This implies that, in order to maintain the low proportion of biallelic expression the transcriptionally inactive periods would have to be even longer than when considering the allele activation dynamics alone. This would make the parameter values in the model even more unrealistic.

Based on this analysis, we conclude that the simple model of independent yet identical transcription activation rate at the two alleles can only reproduce the experimental data under overtly unrealistic conditions.

Stochastic predominant monoallelic expression requires cooperative remodelling of locus transcriptional competence

The shortcoming of the simple model for stochastic transcription activation was the assumption that probability to activate transcription during any time interval is identical for both alleles. Differential chromatin remodeling of the two alleles offers a straightforward solution to this problem as the two alleles can have distinct transcription activation rates. To examine this possibility we extended the model of stochastic gene expression to take into account the dynamics of reversible chromatin remodeling events required for a gene locus to become transcriptionally competent (Figure 4).

We assume that a gene locus can accumulate up to N equivalent (but not independent) reversible chromatin modifications and that the rate at which the locus switches from transcriptionally inactive to active state depends on the n number of modifications it has. The transcription activation rate will be maximal when the locus accumulates $n=N$ modifications, being fully competent, and will be minimal when there are no modifications, $n=0$. Each allelic locus can therefore be in $N+1$ distinct states and undergoes state transitions as it accumulates or loses chromatin modifications. As such, when a locus is in state n , there are n possible ways for it to undergo a reverse remodeling event (lose any of the n accumulated modifications) and $N-n$ ways for it to undergo a forward remodeling event (accumulate one more modification). We assume that forward remodeling events that make the locus more accessible to transcription factors, and the reverse remodeling events that decrease locus competence are competing Poisson processes, and that their respective rates α_n and β_n depend also on the chromatin state n . Thus, the transcription activation rate, the forward and the reverse chromatin state transition rates are functions of n according to equations 4.

$$\begin{aligned} a_n &= a(n+1) \\ \alpha_n &= \alpha(n+\xi) \\ \beta_n &= \beta(N-n+\xi) \end{aligned} \tag{4}$$

where a , α , β and ξ are constants. ξ represents the basal probability of forward or reverse remodeling events and is fixed in $\xi=0.1$ through out this work.

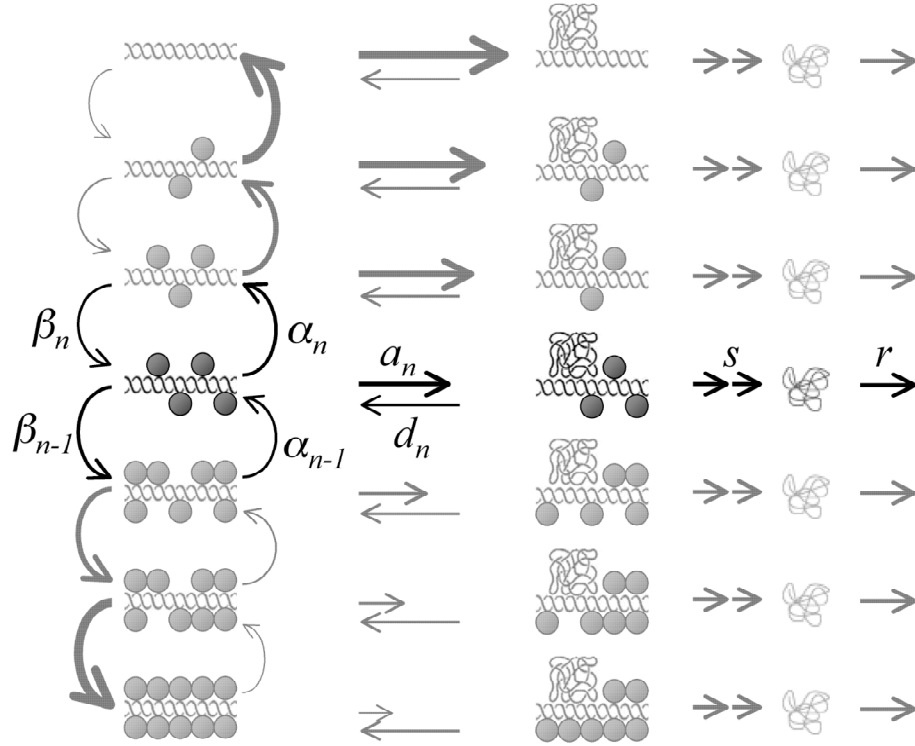


Figure 4- Reversible chromatin modification model of allele expression. This model is an extension to the Cook et al. model in which the rate at which the allele becomes transcriptionally active depends on its chromatin state. Each allele supports up to N equivalent modifications, and can be in $N+1$ distinct states. State transitions happen by forward and reverse modifications, whose rates, respectively a_n and b_n , depend on number of modifications accumulated n . The number of modifications determines the allele competence for transcription through the transcriptional activation rate a_n .

The dynamics of the pattern of allele expression was simulated by Monte Carlo type simulations that, for each independent allele, track the transitions in the number n of chromatin modifications, the switches between transcription inactive and active states, and integrate the differential equation 1 in a piece-wise manner (results illustrated in Figure 5). It should be noticed that model by Cook et al. (Cook et al., 1998) used in previous section, is a special case of latter model from which it can be recovered by setting $n=0$, i.e. transcription activation is independent of the state of the chromatin. This model, although containing many more variables than the previous one, has only three additional parameters n , α , and β .

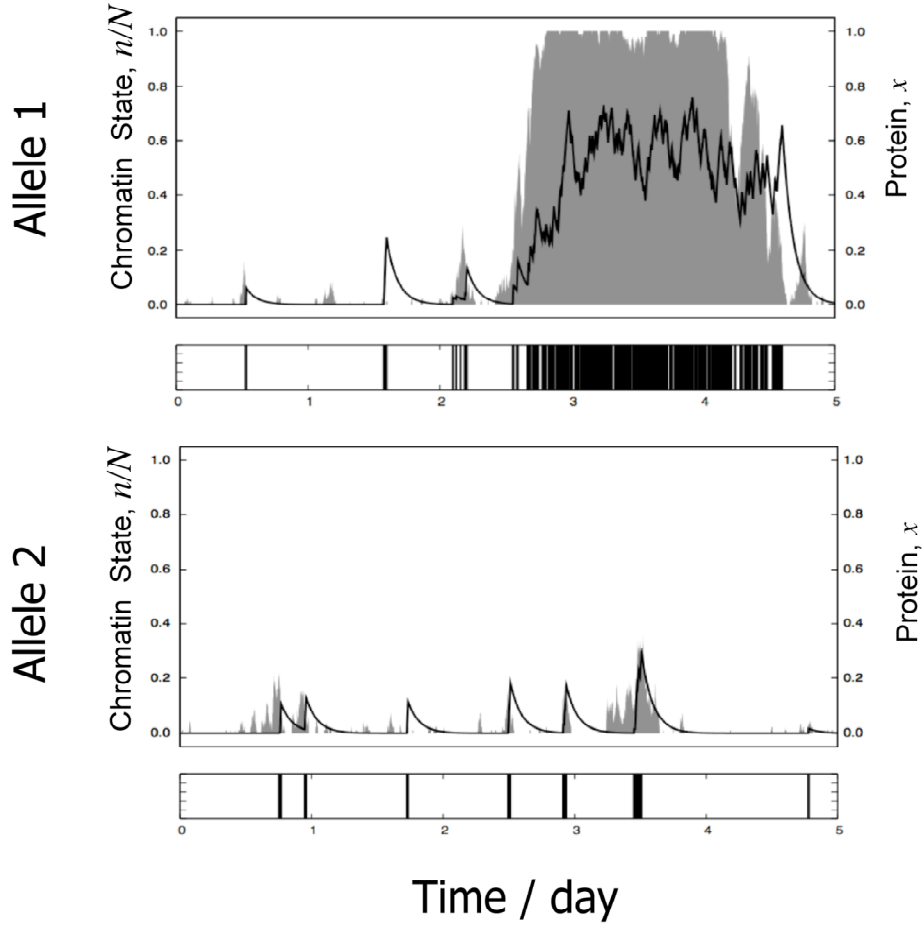


Figure 5 - Representative time series obtained by simulation of the reversible chromatin modification model of allele expression. For each allele the top graph depicts the normalized product concentration (line) and the chromatin competence n/N (grey area). The bottom graph represents the transcriptionally active (black) and inactive (white) states. Parameter values: $N=100$; $a=0.001 \text{ min}^{-1}$; $b=a/0.97 \text{ min}^{-1}$; $\alpha=0.1 \text{ min}^{-1}$; $d=0.1 \text{ min}^{-1}$; $s=0.0083 \text{ au} \cdot \text{min}^{-1}$; $r=0.0083 \text{ min}^{-1}$.

This allele locus competence model can be alternatively formalized as a set of $N+1$ differential equations for the probabilities of an individual allele to be in state n (eqn 5).

$$\begin{cases} \frac{dp_0}{dt} = \beta_1 p_1 - N\alpha_0 p_0 \\ \frac{dp_n}{dt} = (n+1)\beta_{n+1}p_{n+1} + (N-n+1)\alpha_{n-1}p_{n-1} - (n\beta_n + (N-n)\alpha_n)p_n & 0 < n < N \\ \frac{dp_N}{dt} = \alpha_{N-1}p_{N-1} - N\beta_N p_N \end{cases} \quad (5)$$

This formulation is useful since it allows us to get a better quantitative insight into the properties of the mechanism of gene expression. The stationary probability distribution for the locus competence state n is obtained by solving the system of equations 5 for the steady state. Setting the left-hand side to zero, we rewrite the middle general equation in n as:

$$p_{n+1}(n+1)\beta_{n+1} - p_n(N-n)\alpha_n = p_n n\beta_n - p_{n-1}(N-n+1)\alpha_{n-1} \quad (6)$$

Since the left- and right-hand sides of this equality are the same expression in $n+1$ and n respectively, they must be constant and identical. Writing the right-hand side for $n=0$ we conclude that this constant is zero. From this, we get the following recursive expression:

$$p_{n+1} = \frac{(N-n)\alpha_n}{(n+1)\beta_{n+1}} p_n \quad (7)$$

Iterating this expression we obtain all the p_n as a function of p_0 . Since all the p_n must sum 1 we solve the corresponding normalizing condition in order to p_0 obtaining:

$$p_0 = \frac{1}{1 + \sum_{n=1}^N \prod_{i=1}^n \frac{(N-i)\alpha_i}{(i+1)\beta_{i+1}}} \quad (8)$$

Combining equation (8) with the recursive relation (7) we can calculate the stationary probability distribution for all $N+1$ the locus states.

We can then calculate the probability of finding the locus in a transcriptionally active state in steady state as given by equation 9.

$$p_{act} = \sum_{n=0}^N p_n \frac{a_n}{a_n + d} \quad (9)$$

In order to fit the time series for the relaxation kinetics of purified cells we imposed $a_N=d$ and set the ratio a/d in such a way that eqn. 9 gives $p_{act}=p=10\%$. This left one of the rates for chromatin remodeling as a free parameter and we adjusted it so that the whole system reproduces the evolution of the pattern of allele expression of a purified subpopulation that is allowed to relax (Figure 6b). The model is able to describe the experimental data under reasonable parameters values. The remodeling forward and reverse rate happens in a timescale of hours ($\alpha=.001 \text{ min}^{-1}$) and the transcription switching happens at the timescale of minutes ($a=d=0.1 \text{ min}^{-1}$). Although the model is quite flexible on the specific values of these parameters, the ratio between the forward and reverse chromatin remodeling rate constants α/β must be close to unity such that the distribution of number of modifications is bimodal (fig6a).

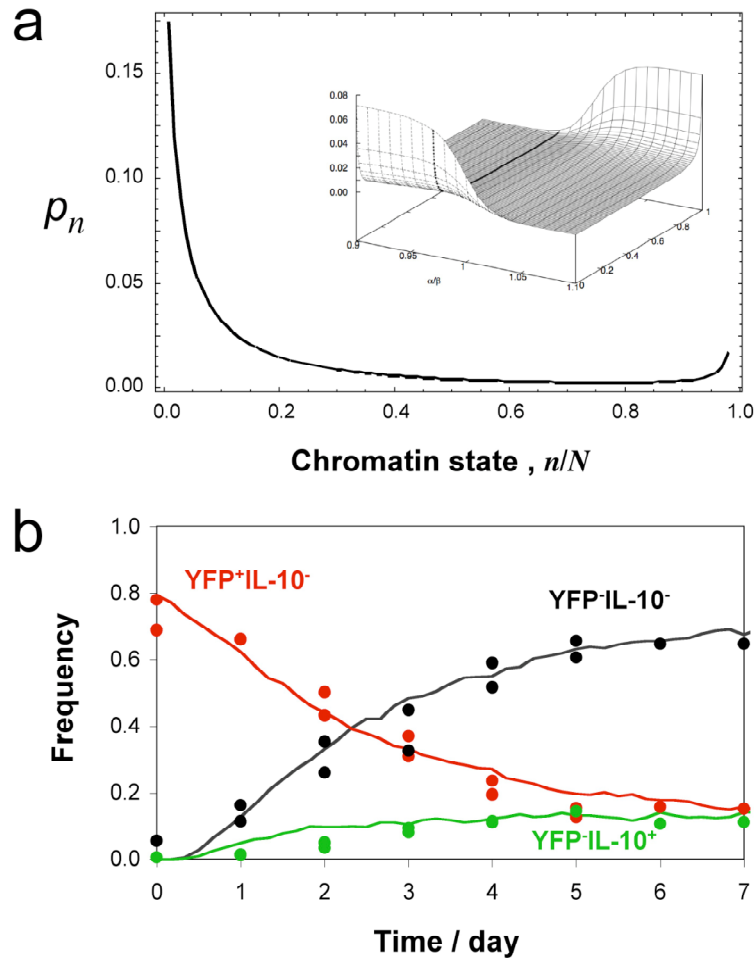


Figure 6- Chromatin states and relaxation kinetics in a cell population in simulations. a) Probability distribution for the chromatin states n . Inset: The same probability distribution for several values of the ratio a/b . b) Time course of the frequency of the indicated subpopulations during the restimulation protocol in experiments (dots) and in simulations (lines). Frequencies were calculated based on 2000 individual cell simulations. In A and B the parameter values were the same as in figure 5.

We conclude that this new model of gene expression, comprising cooperative and reversible chromatin remodeling events that modulate the transcriptional competence of the two allelic locus independently, can easily reproduce the experimental observations on the allele expression pattern and dynamics in a cell populations

Discussion

This article addressed the mechanism underlying the pattern of allelic expression of cytokine genes reporting a new transcription regulation model that takes into account

two concurrent stochastic processes in each allele of a gene: the chromatin remodeling dynamics that changes the transcriptional competence of the locus, and the stochastic switching between transcriptionally active and inactive states of each allele. Besides formalizing in a quantitative way several concepts qualitatively discussed in the immunological literature on cytokine monoallelic expression (Guo et al., 2005; Calado et al., 2006), this new model of transcription regulation captures unique features of eukaryotic transcription that are not often contemplated in stochastic gene regulation models, which are usually designed for prokaryotic systems (Arkin et al., 1998; Banerjee et al., 2004).

In eukaryotic cells, transcription regulation is achieved not only by controlling the concentration of transcription factors and their activity on the gene promoter, but also by enhancer activity and chromatin accessibility at the locus. Eukaryotic cells can afford to regulate the activity of a gene in many ways (Fiering et al., 2000): by adjusting the levels of expression of its products in a “rheostatic” manner, by setting the frequency at which a gene will be expressed in a binary or quantal manner, by defining the proportions of monoallelic or biallelic expression, and by making allele silencing reversible or irreversible. These modes of regulation are naturally accommodated in the model proposed here. Thus, “rheostatic” responses are controlled through the transcriptional activation rate constant a , while the frequency of “quantal” responses can be controlled independently through the ratio α/β , defining a bimodal probability density function of the chromatin states.

The model is also general in the sense that it can be parameterized in such a way as to recover other models in the literature. As mentioned before, when setting the maximal number of modifications promoting transcription to zero, $N=0$, and considering only one allele, one recovers the simplest model of stochastic transcriptional activation that were systematically used to study transcription in prokaryotes (Banerjee et al., 2004). By setting $N=1$, one recovers a simple model describing a minimal two-state chromatin dynamics, as proposed by Blake et al. (Blake et al., 2003). In this respect, it is worth stressing that our model, parameterized with large N , predicts stable bimodal probability distributions of chromatin states that lead naturally to “quantal” responses and stochastic monoallelic expression, not easily obtained in simpler models, without evoking *ad hoc* thresholds. Positive feedback loops play a key role in gene networks as they introduce multistability and potential for cell differentiation (Thomas et al., 1995; Tyson et al., 2003). The reversible chromatin modification model posits that each allele of a gene has a built-in positive feedback loop, stemming from the fact that forward (reverse) modifications increase the chance that other forward (reverse) modifications are

introduced in the locus (eqn. 4). Thus, according to this model, eukaryotic cells would have many more potential states than prokaryotic cells.

Our model did not explicitly take into account cell division or the effect of cell cycle on gene expression and chromatin dynamics. During mitosis chromosomes undergo heavy compactation, involving drastic changes in chromatin structure which are expected to impact on the locus competence for transcription (Chen et al., 2005). The chromatin structure is believed to be restored to its initial condition after mitosis, as this is a prerequisite for lineage identity. But inevitably there must be a lot of noise in this restoration process and the gene locus structure must be somehow robust. The predicted stationary distribution of the chromatin state gives some hints on origins of this robustness. As long as the locus remodeling factors are fixed outside the bimodal region, meaning forward and reverse rate constant are imbalanced, the chromatin state should be highly resilient to perturbations, due to the low probability to reach the opposite state. On the other hand, in genes operating in the bimodal regime (α/β close to 1) these perturbations might in fact induce transitions between the two extreme states of locus competence. In this regime, cell cycle would act as a “catalyst”, accelerating the transitions transcriptionally competent and silenced states. This expectation is in line with the common assertion that cell cycle facilitates the upregulation of inactive genes (Richter et al., 1999)., and suggests that it might actually be involved in accelerating the rate at which a cell population reaches a stationary pattern of predominant monoallelic expression (Richter et al., 1999). This suggests that a quantitative difference in the absolute values of forward and reverse modification rates (measured by the parameters α and β respectively) may underly the qualitative difference between those instances of strict stochastic monoallelic expression, which are stable in the single cell across the cell cycle, such as odorant receptors genes and *Tlr4* gene, and those instances of predominant stochastic monoallelic expression of cytokine genes, which are unstable at the single cell level.

Imprinted genes are epigenetically marked in haploid gametes and this structure is maintained in the fertilized egg and inherited by the somatic cells. However, in all the other instances of stable monoallelic expression, such as X-chromosome inactivation or allelic exclusion, diploid cells have to choose between two virtually identical alleles. The solution to the problem seems to be making a random choice first, and then make this choice irreversible and inheritable by the daughter cells. The reversible chromatin remodeling model offers two stochastic processes — changing between chromatin states and switching between transcriptionally active and inactive states — that could be explored by the cell to make a random allele choice. According to the parameters that fit the observations

on the *Il10* locus expression, the two stochastic processes have different time scales. Switching between transcriptionally active and inactive states happens within minutes to hours, while switching between fully competent and incompetent chromatin states happens at a time scale that affords longer periods of monoallelic expression. Considering these time scales, the most likely process for the random allele choice will involve chromatin remodeling dynamics since this offers more time. More time that is necessary for the “quality control” of the chosen allele, and for activation of an irreversible process, such as shutting down the rag-genes in allele exclusion of antigen-receptor genes or silencing definitively all odorant receptor genes and alleles in olfactory neurons. In fact, theoretical calculations on the frequency of post-recombination γ locus configurations in $\gamma\delta$ thymocytes, emphasized the importance of a period of monoallelic expression for efficient allelic and isotype exclusion (Boucontet et al., 2005; Sepulveda et al., 2005). Thus, although the mechanism of cytokine predominant monoallelic expression at first sight looks as an outcast of the family of more strict monoallelic expressed genes, it might actually play a key role in the allele choice of the more complex mechanisms controlling allelic exclusion, X-chromosome inactivation, or TLR-4 expression. Genes in which active allele choice is deterministic such as parental imprinted genes, where one of the alleles is silent and the other inducible, call into action features that are not contemplated in the model but easily incorporated, since all they imply is that the forward and reverse modification rates are distinct for the two alleles.

Predominant monoallelic expression arises so naturally in our model of eukaryotic expression that it is fair to ask: How many genes are expressed in predominantly monoallelic or biallelic manner? If the transcription of the two alleles is independently and controlled as we proposed, the probability of the joint event of activating simultaneously both alleles is low. More precisely monoallelic expression will be as frequent as biallelic expression when $2p(1-p)=p^2$ i.e. $p=2/3$, and monoallelic expression will predominate when $p<2/3$. Under these conditions the expected frequency of expressing cells would be $1-(1-p)(1-p)$ i.e. 88.(8)%. It is not uncommon to have less than 90% of the cells expressing a gene following its induction. This suggests that predominant monoallelic expression might be a common phenomenon, not just restricted to cytokine genes. Moreover, we expect that even in a cell population that contains mostly biallelic expressors, a subpopulation of monoallelically expressing cells will be also present.

This leads us to the question: why are the cytokine genes regulated to be in the regime of low probability of expression, where the frequency of biallelic expressing cells is lower than 5%? Cytokines are produced towards the synapses between

cooperating cells, where a burst of exocytosis saturates cytokine receptors (Coombs and Goldstein, 2004), leading to full response of the target cell. Under these conditions, the biological function, which is quantified by the number cells responding to the cytokine, must be controlled via the frequency of cytokine producing cells rather than via the amount of cytokine produced by all cells, as discussed before (Guo et al., 2005). Also, many of these cytokines are involved in autoregulatory loops that induce their own expression in other cells, either directly (e.g. IL-2 or IL-4) or indirectly (e.g. IFN- γ), and therefore, keeping low frequency of cells expressing these “cell fate” controlling cytokines ensures the persistence of precursors.

Finally, we return to the issue of the putative selective advantage that predominant monoallelic expression of some genes might give to diploid organisms. In multicellular organisms, despite a gene being stochastically monoallelic expressed, the selective advantage of diploidy is still present since the organism, as a whole, does express the two alleles, although in different cell subpopulations. Thus, monoallelic expression at the single cell level will be neutral provided that the two-fold reduction in the number of gene expressing cells is not physiologically relevant, or if the number of these cells is itself regulated by some population feedback mechanism. In this way, stochastic monoallelic expression will not only mask the effects of recessive deleterious mutations as expected, but it will also minimize the deleterious effects of carrying an allele that behaves as dominant negative in single cells (Schimenti, 2001; Pereira et al., 2003). Notwithstanding this potential adaptive advantage, the model developed here suggests that stochastic predominant monoallelic expression, like many other phenotypes (Gould and Lewontin, 1979), did not evolve under direct action of natural selection. Instead, we favor the view (Guo et al., 2005; Calado et al., 2006) according to which the phenotype under selection is the low frequency of gene expressing cells, which given the mechanism of gene expression regulation in eukaryotes, leads to predominant monoallelic expression.

Acknowledgements

The authors are grateful to Andreas Radbruch, Jocelyne Demengeot, and very specially Matthias Haury for many useful discussions that culminated in this work. Carvalho made his contribution to this work during an internship as a student of the

Gulbenkian PhD Programme in Biomedicine. Calado and Paixão acknowledge the financial support of Fundação para a Ciência e a Tecnologia (fellowships SFRH/BD/2973/2000 and SFRH/BD/10550/2002, respectively).

References

- Affolter, M., A. Percival-Smith, et al. (1990). "DNA binding properties of the purified Antennapedia homeodomain." Proc Natl Acad Sci U S A **87**(11): 4093-7.
- Arkin, A., J. Ross and H. H. McAdams (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." Genetics **149**(4): 1633-48.
- Banerjee, B., S. Balasubramanian, et al. (2004). "Tracking operator state fluctuations in gene expression in single cells." Biophys J **86**(5): 3052-9.
- Bix, M. and R. M. Locksley (1998). "Independent and epigenetic regulation of the interleukin-4 alleles in CD4+ T cells." Science **281**(5381): 1352-4.
- Blake, W. J., K. A. M, C. R. Cantor and J. J. Collins (2003). "Noise in eukaryotic gene expression." Nature **422**(6932): 633-7.
- Boucontet, L., N. Sepulveda, J. Carneiro and P. Pereira (2005). "Mechanisms controlling termination of V-J recombination at the TCRgamma locus: implications for allelic and isotypic exclusion of TCRgamma chains." J Immunol **174**(7): 3912-9.
- Calado, D. P., T. Paixao, D. Holmberg and M. Haury (2006). "Stochastic monoallelic expression of IL-10 in T cells." J Immunol **177**(8): 5358-64.
- Carneiro, J., T. Paixão, et al. (2005). "Immunological Self-Tolerance: Lessons from Mathematical Modeling." J Comp Appl Math **184**: 77-100.
- Chen, D., M. Dundr, et al. (2005). "Condensed mitotic chromatin is accessible to transcription factors and chromatin structural proteins." J Cell Biol **168**(1): 41-54.
- Cook, D. L., A. N. Gerber and S. J. Tapscott (1998). "Modeling stochastic gene expression: implications for haploinsufficiency." Proc Natl Acad Sci U S A **95**(26): 15641-6.
- Coombs, D. and B. Goldstein (2004). "Effects of the geometry of the immunological synapse on the delivery of effector molecules." Biophys J **87**(4): 2215-20.
- Day, T. and R. Bonduriansky (2004). "Intralocus sexual conflict can drive the evolution of genomic imprinting." Genetics **167**(4): 1537-46.

- Ekker, S. C., K. E. Young, D. P. von Kessler and P. A. Beachy (1991). "Optimal DNA sequence recognition by the Ultrabithorax homeodomain of *Drosophila*." Embo J **10**(5): 1179-86.
- Fiering, S., E. Whitelaw and D. I. Martin (2000). "To be or not to be active: the stochastic nature of enhancer action." Bioessays **22**(4): 381-7.
- Gimelbrant, A. A. and A. Chess (2006). "An epigenetic state associated with areas of gene duplication." Genome Res **16**(6): 723-9.
- Goldmit, M. and Y. Bergman (2004). "Monoallelic gene expression: a repertoire of recurrent themes." Immunol Rev **200**: 197-214.
- Gould, S. J. and R. C. Lewontin (1979). "The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme." Proc R Soc Lond B Biol Sci **205**(1161): 581-98.
- Guo, L., J. Hu-Li and W. E. Paul (2004). "Probabilistic regulation of IL-4 production in Th2 cells: accessibility at the *Il4* locus." Immunity **20**(2): 193-203.
- Guo, L., J. Hu-Li and W. E. Paul (2005). "Probabilistic regulation in TH2 cells accounts for monoallelic expression of IL-4 and IL-13." Immunity **23**(1): 89-99.
- Heard, E. and C. M. Disteche (2006). "Dosage compensation in mammals: fine-tuning the expression of the X chromosome." Genes Dev **20**(14): 1848-67.
- Held, W., J. Roland and D. H. Raulet (1995). "Allelic exclusion of Ly49-family genes encoding class I MHC-specific receptors on NK cells." Nature **376**(6538): 355-8.
- Hollander, G. A., S. Zuklys, et al. (1998). "Monoallelic expression of the interleukin-2 locus." Science **279**(5359): 2118-21.
- Hu-Li, J., C. Pannetier, et al. (2001). "Regulation of expression of IL-4 alleles: analysis using a chimeric GFP/IL-4 gene." Immunity **14**(1): 1-11.
- Kelly, B. L. and R. M. Locksley (2000). "Coordinate regulation of the IL-4, IL-13, and IL-5 cytokine cluster in Th2 clones revealed by allelic expression patterns." J Immunol **165**(6): 2982-6.
- Kuznetsov, V. A., G. D. Knott and R. F. Bonner (2002). "General statistics of stochastic process of gene expression in eukaryotic cells." Genetics **161**(3): 1321-32.
- Kwon, H., S. Park, et al. (2001). "Determination of binding constant of transcription factor AP-1 and DNA. Application of inhibitors." Eur J Biochem **268**(3): 565-72.
- Otto, S. P. and D. B. Goldstein (1992). "Recombination and the evolution of diploidy." Genetics **131**(3): 745-51.

- Pereira, J. P., R. Girard, et al. (2003). "Monoallelic expression of the murine gene encoding Toll-like receptor 4." Nat Immunol **4**(5): 464-70.
- Richter, A., M. Lohning and A. Radbruch (1999). "Instruction for cytokine expression in T helper lymphocytes in relation to proliferation and cell cycle progression." J Exp Med **190**(10): 1439-50.
- Riviere, I., M. J. Sunshine and D. R. Littman (1998). "Regulation of IL-4 expression by activation of individual alleles." Immunity **9**(2): 217-28.
- Schimenti, J. (2001). "Monoallelic gene expression in mice: who? When? How? Why?" Genome Res **11**(11): 1799-800.
- Sepulveda, N., L. Boucontet, P. Pereira and J. Carneiro (2005). "Stochastic modeling of T cell receptor gamma gene rearrangement." J Theor Biol **234**(2): 153-65.
- Serizawa, S., K. Miyamichi and H. Sakano (2004). "One neuron-one receptor rule in the mouse olfactory system." Trends Genet **20**(12): 648-53.
- Singh, N., F. A. Ebrahimi, et al. (2003). "Coordination of the random asynchronous replication of autosomal loci." Nat Genet **33**(3): 339-41.
- Spicuglia, S., D. M. Franchini and P. Ferrier (2006). "Regulation of V(D)J recombination." Curr Opin Immunol **18**(2): 158-63.
- Thomas, R., D. Thieffry and M. Kaufman (1995). "Dynamical behaviour of biological regulatory networks--I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state." Bull Math Biol **57**(2): 247-76.
- Trieu, M., J. M. Rhee, N. Fedtsova and E. E. Turner (1999). "Autoregulatory sequences are revealed by complex stability screening of the mouse brn-3.0 locus." J Neurosci **19**(15): 6549-58.
- Tyson, J. J., K. C. Chen and B. Novak (2003). "Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell." Curr Opin Cell Biol **15**(2): 221-31.
- Wolf, J. B. and R. Hager (2006). "A Maternal-Offspring Coadaptation Theory for the Evolution of Genomic Imprinting." PLoS Biol **4**(12): e380.
- Yao, J., K. M. Munson, W. W. Webb and J. T. Lis (2006). "Dynamics of heat shock factor association with native gene loci in living cells." Nature **442**(7106): 1050-3.

IV

**Implications for cell signaling and
population responses**

The Impact of Heterogeneity In Common Network Motifs

Tiago Paixão* and Jorge Carneiro*

**Instituto Gulbenkian de Ciência*

Summary

Modularity in protein or gene interaction networks holds the promise for an understanding of the overall network dynamics using simple abstractions of these “modules”. Although simple, these small networks are often nonlinear in nature and thus are usually characterized through the use computational models describing bulk cell population responses to perturbations. This means that these models are attempting to reproduce the mean behavior of the cell, in the hope that this is the same as the average cell.

However, recent observations show that cell populations display a remarkable heterogeneity in their transcript or protein copy numbers. Since these networks often display nonlinear behavior it is not warranted that the average response of the population is identical to the response of an average cell or in fact to any cell in the population.

Here we evaluate the impact of variability in protein copy numbers in increasingly complex, yet small, networks. We study how a population that shows variability in copy number in key proteins responds, both in steady state as dynamically, to stimulus for each of these modules. We show that population bulk responses can be misleading in inferring the network within the single cell. Moreover, we show that model inference under these conditions is further impaired due to the inability of the mechanism that generates such responses (at the single cell level) to reproduce the mean behavior of the population.

Introduction

The cellular metabolism and physiology is organized in complex networks of interactions between proteins and genes whose statistical and dynamical properties are impossible to understand without the aid of mathematical tools (Kitano, 2002 ; Alon, 2007). Recent years have seen an overwhelming amount of work concerning this subject, resulting in the emergence of a new field, which came to be known as systems biology (Kitano, 2002). One major result that ensued from such works was the observation that these networks presented a clustered (Han et al., 2004) topology and could be separated in smaller networks, hence facilitating their analysis. The mathematical modeling of these small networks ("modules") could hopefully allow the understanding of the whole network, in the same way as an engineer would understand an electronics blueprint without needing to know how a transistor functions internally.

Mathematical cell biology has been pursuing this for a long time and many mathematical models have been proposed and calibrated by fitting to experimental data. Typically, these models are formulated as mean field models, which means that it is assumed that the average behavior of the population is identical to the behavior of the average cell. These models are then fitted to data averaged from bulk cell responses from populations of cells, such as kinetic data from microarrays (Kholodenko et al., 2002).

However, recent observations suggest that stochastic effects play a major role in the cells metabolism and that there is high variability in the amount of protein expressed by each cell, even in clonal populations (Elowitz et al., 2002; Raser and O'Shea, 2005; Sigal et al., 2006). Moreover it was also observed that the shape of the distribution of specific proteins is often nonsymmetrical, as discussed in chapter II and in (Furusawa et al., 2005; Krishna et al., 2005). These observations, coupled to the ubiquitous nonlinearity in biochemical processes, pose a difficulty for the aforementioned assumption. Since it is not granted that, under general conditions, the population's average is followed by any cell, the parameters estimated might be inducing wrong interpretations of the underlying cellular mechanism.

In a distributed population, each cell will have its own representation of the same stimulus, creating a distributed response. These distributed responses are then fed to other mechanisms creating the possibility that huge amplifications of variance take place.

It is then imperative to describe the populations response under heterogeneity conditions; how is the image of a stimulus propagated through each mechanism; how is the response distribution across a population and under which conditions is legitimate to equate the average response of the population with the response of the average cell. Moreover, a comprehensive study of how each module transforms its input might offer the possibility of identifying underlying mechanisms by characterizing the distributions of intervenient elements and their respective output.

This work tries to place us a step forward towards these goals by analyzing typical molecular mechanisms under heterogenous conditions, characterizing how each mechanism transforms an input, both in steady state and in dynamical conditions, and comparing the average behavior of the population with the response of the average cell.

Results

For any given response function and distribution of its underlying constituents ($\rho(p)$), the population's average response $R(s;p)$ is only equal to the response of the average cell if the individual response function $r(s;p)$ is linear:

$$R(S) = \int r(s;p)\rho(p)dp = \int (a(S)p + b(S))\rho(p)dp = a(S) \langle p \rangle + b(S)$$

Even under these conditions where linearity is satisfied, the shape of the response distribution depends on the specific shape of the distribution of protein content of the population. In order to give a somewhat comprehensive overview of the type of behavior we can expect from a distributed population we analyzed most of mechanisms that featured in the comprehensive review by (Tyson et al., 2003). In what follows, we illustrate the effects to the structure of some of the modules in conditions where the population is distributed and pinpoint the differences between the average of such population and the average cell.

Hyperbolic responses

The simplest example of one such “module” is the simple mechanism depicted in Figure 1 where a pool of proteins (r) is reversibly modified by some external signal (S) at a constant rate. This process is truly ubiquitous and models cellular processes such as phosphorylation / dephosphorylation loops, receptor binding by non-limiting

ligand and any kind of bimolecular reversible reaction in which one of the reactants is not limiting.

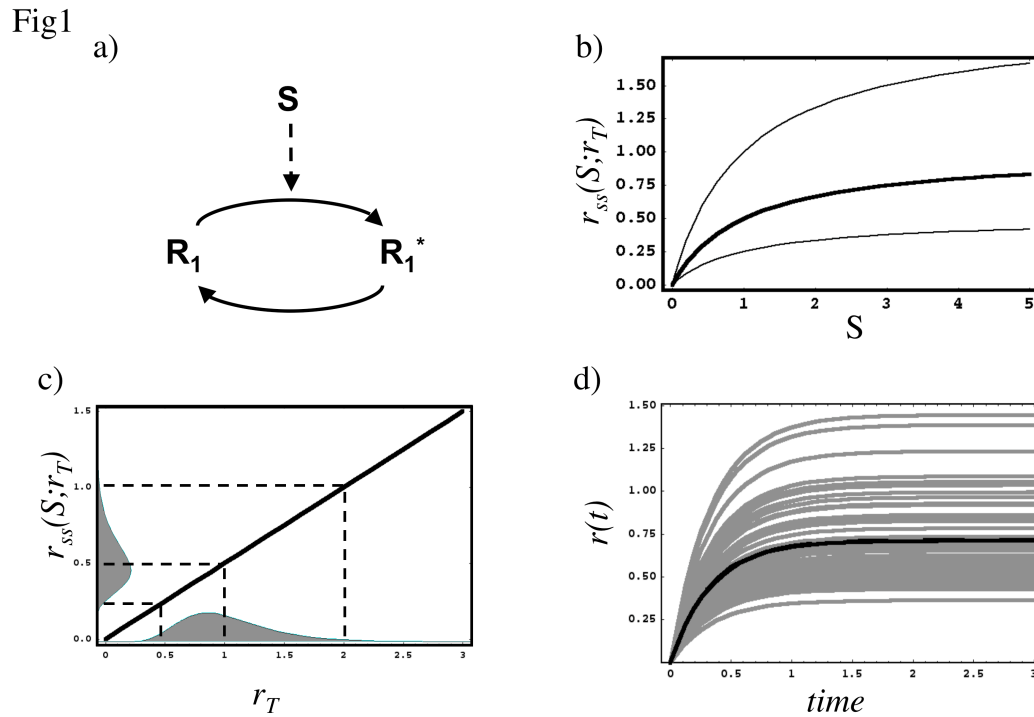


Figure 1: The basic hyperbolic mechanism. a) diagram of the mechanism. b) Steady state as a function of signal strength. Thick line represents the steady state for a nominal total protein amount ($r_T=1$) and thinner lines represent 2-fold increase and decrease of that value. c) thick line represents the steady state as a function of total protein amount (for $S=1$). The horizontal histogram represents the distribution of r_T and the vertical histogram shows how that is mapped in terms of steady state output by this module. d) several kinetic trajectories for a population with r_T distributed as a lognormal distribution with parameters $m=\log(1)$ and $s=0.25$. Continuous black line represents the average cell of this population and dashed line (overlapping) the average of all trajectories. For all figures $k_1=1$ and $k_2=1$.

This process has the advantage of being exactly solvable and we'll use it as an example for many of the procedures we'll use in order to investigate the role of heterogeneity in population responses.

This process can be described by the following rate equation

$$\frac{dr(t)}{dt} = k_1 S(r_T - r(t)) - k_2 r(t)$$

the exact solution for this differential equation is:

$$r(t) = (r(0) - \frac{k_1 S}{k_1 S + k_2} r_T) e^{-(k_1 S + k_2)t} + \frac{k_1 S}{k_1 S + k_2} r_T$$

As can easily be seen, the steady state for this equation is $r_{ss} = \frac{k_1 S}{k_1 S + k_2} r_T$. Now let's consider that we have a population that displays heterogeneity in the total amount of this protein (r_T). This means that the population response will be the superposition of the different responses of each cell.

As can be seen in Figure 1, both the steady state response and the dynamics are just a constant multiplied by the total amount of protein in each cell, the distributed parameter, i.e, they are linear with respect to the distributed parameter. As such, the average response of the population is just the response of the average cell; the average is preserved under the population representation of the stimulus. In fact, even the shape of the lognormal distribution is preserved (Figure 1c), this operation affecting only the scale parameter. Notice however that both the mean value and the variance of the output are necessarily smaller than the original one, since $\frac{k_1 S}{k_1 S + k_2} < 1$. However, because the average and the standard deviation are affected by the same value (under a linear transformation), the variation coefficient is kept constant.

However, when we couple two of these modules together (as commonly seen in phosphorylation cascades) the situation changes, as the dynamics of the second module are indeed affected by the output of the first. If we assume that the first module reaches equilibrium faster than the second, this system becomes equivalent to having only one of such modules and subjecting it to a distributed input. This analogy makes the system much easier to analyze and we'll exploit that in the next few paragraphs.

First of all, the steady state response of this compound system, $r_2^{ss} = \frac{k_{11} k_{21} r_{1T} S}{k_{12} k_{22} + k_{11} k_{22} S + k_{11} k_{21} r_{1T} S} r_{2T}$, is nonlinear in the total concentration of the first module.

Therefore, the average steady state value of the population is, in principle, different from the steady state of the average cell. The question is how different they are, i.e., if the average of the population is a good approximation to the average cell. When most of the cells have $r_{1T} \ll \frac{k_{12} k_{22} S + k_{12} k_{22}}{k_{11} k_{21}}$ the system behaves linearly with r_{1T} , and the output distribution will be the product of r_{1T} and r_{2T} . If the total concentration of r_1 and r_2 are distributed as lognormal distributions then this product will also be a lognormal distribution with parameters ($m' = m_1 + m_2; s'^2 = s_1^2 + s_2^2$), i.e., the variance (in logscale) of the output distribution will be the sum of the variances of the two modules, resulting in a bigger dispersion of the output values. On the other hand, if most cells have $r_{1T} \gg \frac{k_{12} k_{22} S + k_{12} k_{22}}{k_{11} k_{21}}$, then the output of the first module is completely saturated and the population will essentially display the distribution of the second module. The shape of

the distribution for intermediate values of r_{1T} depends on the specific shape of the distributions of r_{1T} and r_{2T} .

In terms of kinetic behavior, the average of the population, because of the nonlinearity of the kinetic solution on the distributed parameter, is in general different from the single cell solution.

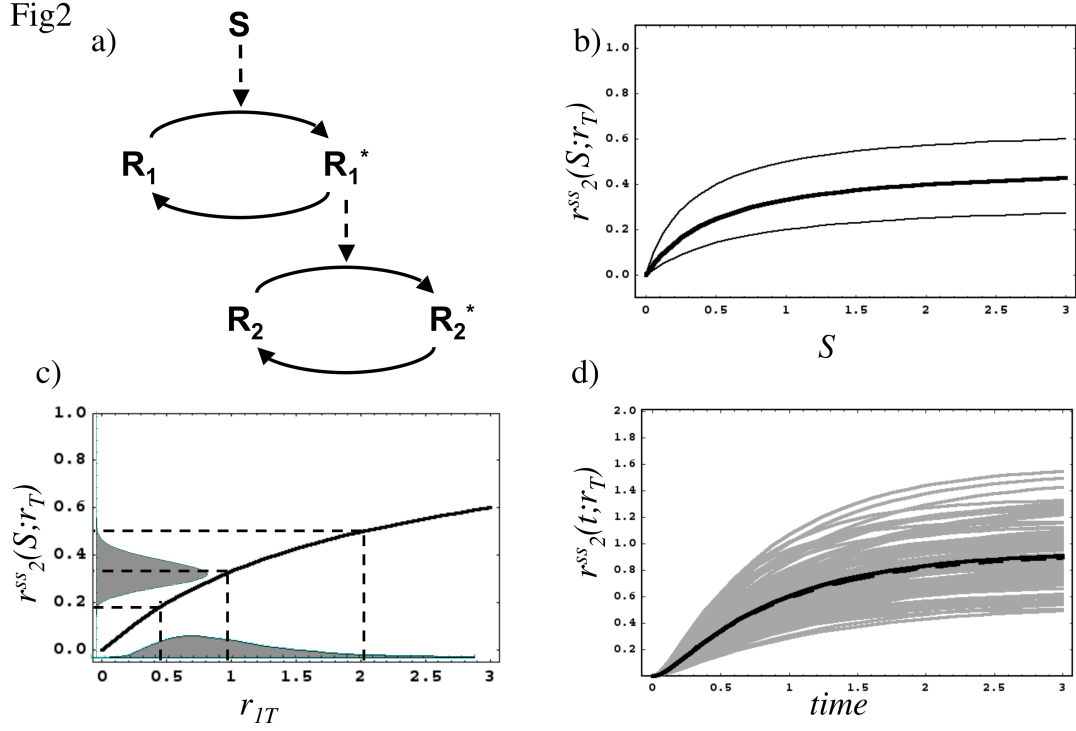


Figure 2: Two hyperbolic modules coupled. a) diagram of the mechanism. b) Steady state as a function of signal strength. Thick line represents the steady state for a nominal total protein amount ($r_{1T}=1$) and thinner lines represent 2-fold increase and decrease of that value c) thick line represents the steady state as a function of total protein amount (for $S=1$). The horizontal histogram represents the distribution of r_T and the vertical histogram shows how that is mapped in terms of steady state output by this module. Notice how the shape of the output distribution is contracted by the nonlinear relationship of r_2^{ss} and r_{2T} . d) several kinetic trajectories for a population with r_{1T} and r_{2T} distributed as a lognormal distribution with parameters $M=\log(1)$ and $S=0.25$. Continuous black line represents the average cell of this population and dashed line the average of all trajectories. For all figures $k_{11}=k_{12}=k_{21}=k_{22}=1$.

In fact, by assuming, for mathematical simplicity, that the first module is saturated and faster than the second module, we reduce the problem to the dynamics of the second module with a distributed input. If we further assume that the underlying distribution of the input is a gamma distribution we are able to compute the average dynamics of the population:

$$\langle r(t) \rangle = \int r(t; r_T) \gamma(r_T; \alpha, \theta) dr_T = 1 - \theta^{-k} \left(\frac{1}{\theta} + t \right)^{-k}$$

which is markedly different from the single cell response. In fact, if the variance of such distributions is not very small there will be appreciable differences between the average response of the population and the response of the average cell. However, for reasonable variances of the underlying distribution, the average of the population does not deviate significantly from the behavior of the average cell (Figure.2d).

Sigmoidal responses

The previous mechanism always gives very gradual increments of the response as function of the signal and it's one of the simplest mechanisms that the cell uses in order to transduce signals. A plausible mechanism was theoretically predicted (Goldbeter and Koshland, 1981) that, due to the saturation of the phosphorylation/dephosphorylation processes, induces a more steep response curve, a phenomenon known as zero order ultrasensitivity.

This model is described by:

$$\frac{dr}{dt} = \frac{k_1 S (r_T - r)}{K_{m1} + r_T - r} - \frac{k_2 r}{K_{m2} + r}$$

the steady state solution is the given by the Goldbeter-Koshland function, defined as:

$$r^{ss} = G(k_1 S, k_2, \frac{K_{m1}}{r_T}, \frac{K_{m2}}{r_T}) r_T = \frac{2k_1 k_2 S r_T}{k_2 - k_1 S + k_2 \frac{K_{m1}}{r_T} + k_1 S \frac{K_{m2}}{r_T} + \sqrt{(k_2 - k_1 S + k_2 \frac{K_{m1}}{r_T} + k_1 S \frac{K_{m2}}{r_T})^2 - 4(k_2 - k_1 S) k_1 S \frac{K_{m2}}{r_T}}}$$

As can easily be derived, the threshold signal is given by $S = \frac{k_1 (2K_{m1} + r_T)}{k_2 (2K_{m2} + r_T)}$. So, if $K_{m1} = K_{m2}$, the total concentration of r has no influence on the critical signal that induces the transition.

Fig 3.

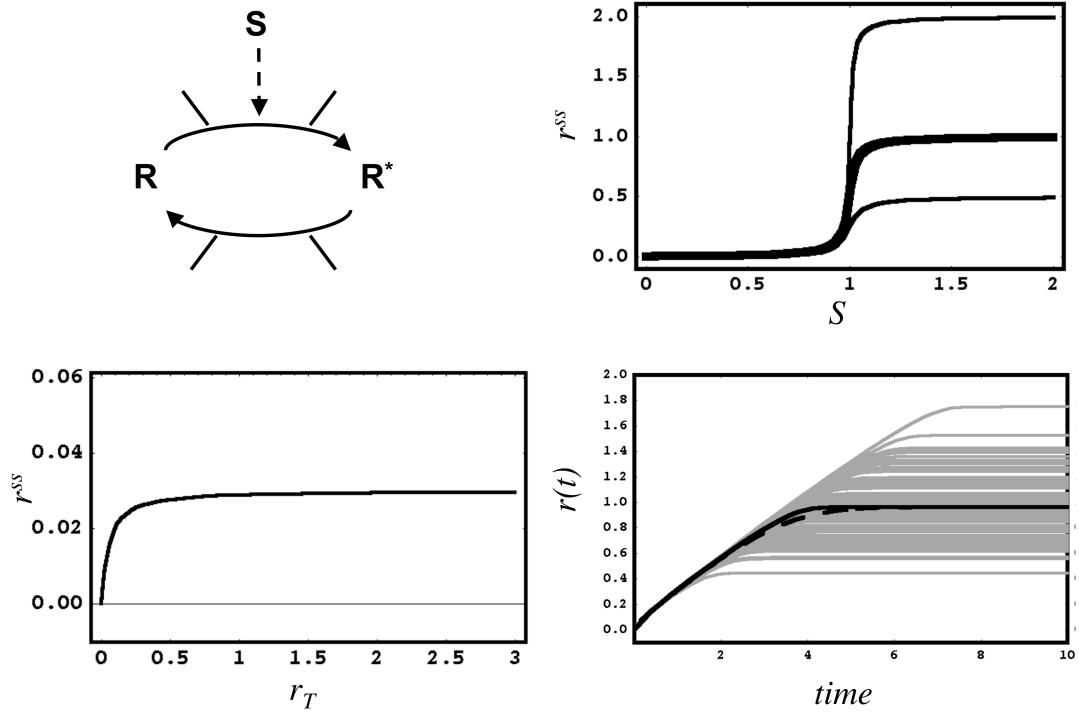


Figure 3: The Koshland-Golbeter switch. a) diagram of the mechanism. b) Steady state as a function of signal strength. Thick line represents the steady state for a nominal total protein amount ($r_{1T}=1$) and thinner lines represent 2-fold increase and decrease of that value. c) thick line represents the steady state as a function of total protein amount (for $S=.75$). This steady state would map all the variance of population to a point. The steady state above the critical signal ($S>1$) is essentially a straight line and the same conclusions from the simple hyperbolic module apply. d) several kinetic trajectories for a population with r_T distributed as a lognormal distribution with parameters $m=\log(1)$ and $s=0.25$. Continuous black line represents the average cell of this population and dashed line (overlapping) the average of all trajectories. For all figures $k_1=k_2=1$ and $K_{m1}=K_{m2}=0.01$.

However, if this condition is not satisfied each cell will have its own critical signal intensity beyond which it would maximally respond. This means that by titrating the signal more and more cells respond, giving rise to a gradual steady state response curve, which does not reflect the reality of any single cell. As such, a single input strength can give rise to a bimodal distribution of outputs, making for a system able to regulate the number of responding cells. This might be useful for an organism when there is a need to differentiate cells while maintaining a pool of progenitors. However, in order to have switch like behavior, K_{m1}/r_T and K_{m2}/r_T should be much less than 1. Under these conditions, the variance of the threshold distribution is always much less than the one from the underlying distribution (the same considerations for

the propagation of variance made above about the 2 hyperbolic modules applies here). However, by coupling this module to a hyperbolic module (Figure.4), effectively providing a distributed input to this module, we recover the bimodal output distribution, as if the population were indeed distributed in the critical signal values. These bimodal distributions are typically the hallmark of feedback loops, both positive and negative, as has been shown by several authors (Yang and Iglesias, 2006; Ingolia and Murray, 2007).

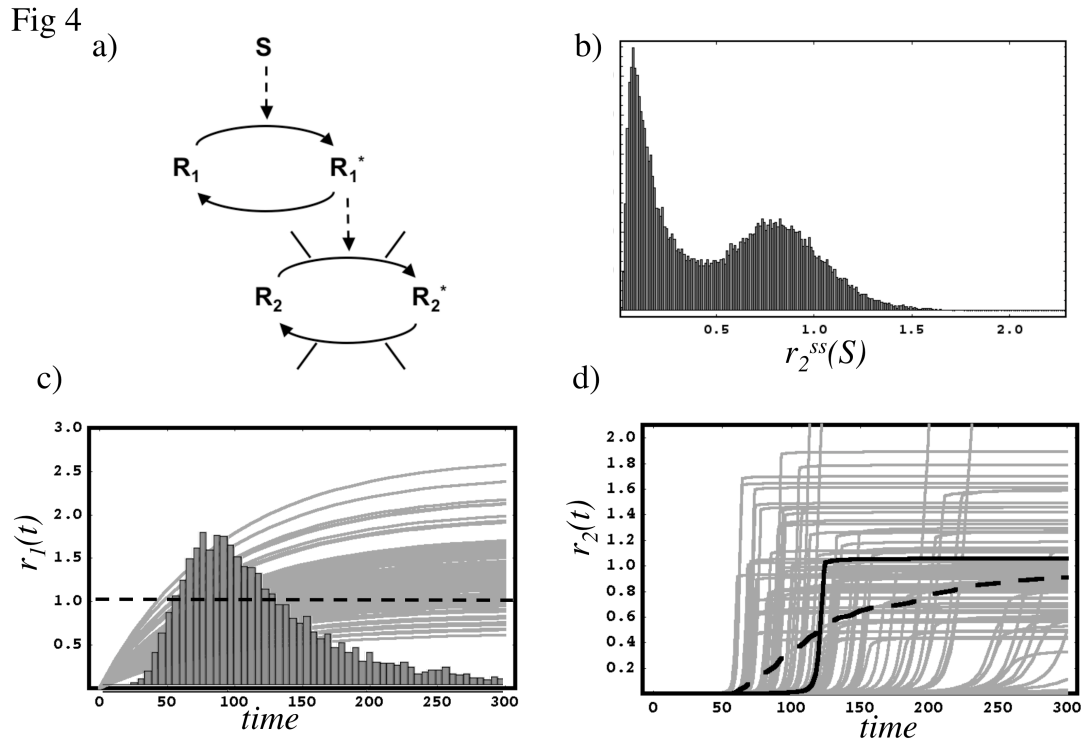


Figure 4: A hyperbolic module couple to a Koshland-Goldbeter switch. a) diagram of the mechanism. b) output distribution of the steady states for signal strength $S=1$. c) Several kinetic trajectories for the hyperbolic module. The threshold value for the switch is the dashed line. Histogram is the distribution for times to reach the threshold value for a population with r_{1T} distributed as a lognormal distribution with parameters $m=\log(1)$ and $s=0.25$. d) Several kinetic trajectories for the same population. Continuous black line represents the average cell of this population and dashed line the average of all trajectories. Notice that there are no cells that follow the dynamics of the average of the population. For all figures $k_{11}=0.001$, $k_{12}=0.01$, $k_1=k_2=10$, $K_{m1}=0.001$, $K_{m2}=0.001$.

Kinetically, this last situation also raises several issues, since as the upstream component accumulates signal to different steady states due to its intrinsic heterogeneity, it reaches the downstream threshold at wildly different time points.

The time each cell takes to reach the critical value is $t_{crit} = \frac{1}{k_{11}S + k_{12}} \log\left(\frac{r_1^{ss}}{r_1^{ss} - r_1^{crit}}\right)$.

From this expression we can derive the distribution of these times, given a distribution for the steady state values r_1^{ss} . Since each cell reaches the threshold at a different time and when it reaches it makes a swift transition to the high steady state (Figure 4), the average kinetics of the population will be a very graded curve to which could be easily be fitted a Michaelis-Menten type kinetics, completely discarding the discrete nature of the dynamics at the single cell level. In fact, in order to fit the model to these dynamics, one would estimate parameters that would place the model in a region where there was no switch like behavior, behaving instead as typical michaelis-menten dynamics (or the two hyperbolic modules of the previous section), thereby rejecting the zero-order ultrasensitivity model, that produced the synthetic data.

Feedback and bifurcation induced switches

In the previous section, we wired a hyperbolic module to a Koshland-Goldbeter module and achieved bimodal steady state distributions solely due to the heterogeneity within the hyperbolic module. However, such circuitry is still what Tyson *et al* call a buzzer; it falls back to the inactive state as soon as the signal is turned off. Any irreversible switch requires a bistable system that exhibits the coexistence of two stable steady states and has the input regulate the transition from one to the other. This is typically achieved by feedback or feedforward loops within the regulatory system. This type of module is characterized by a critical signal intensity beyond which the irreversible switch takes place and it's of paramount importance to determine how this behavior is affected by the observed heterogeneity in cell populations.

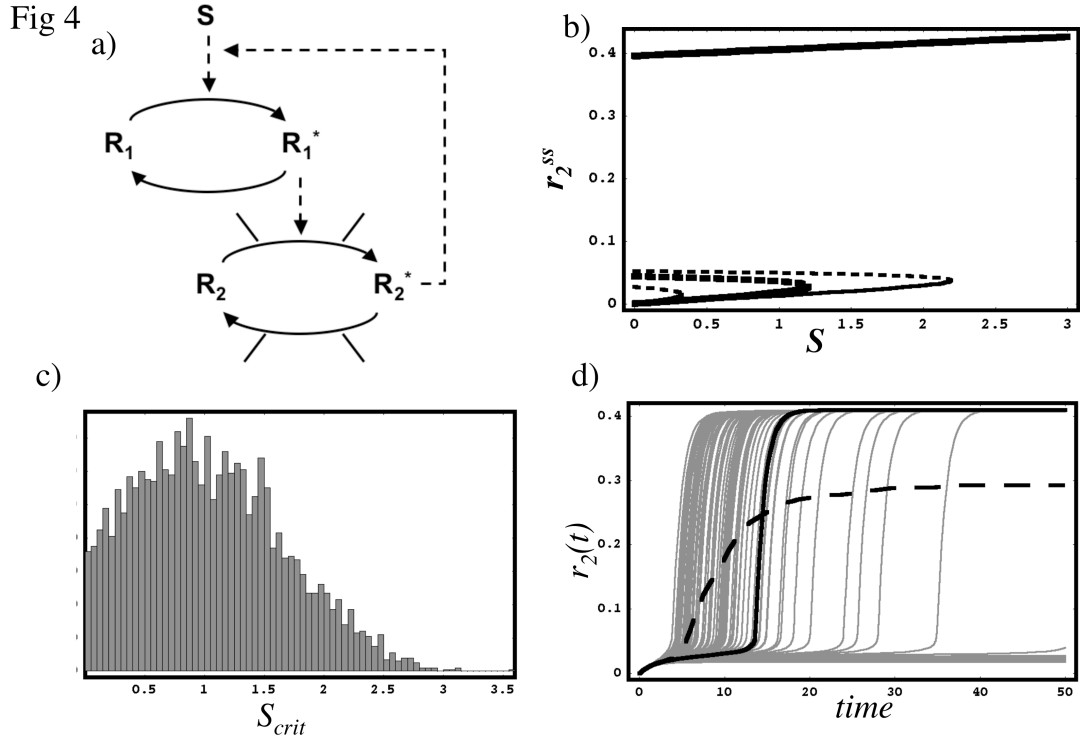


Figure 5: A mutual activation module. a) diagram of the mechanism. b) Steady state as a function of signal strength. Thick continuous line represents the steady state for a nominal total protein amount ($r_{1T}=1$) and thinner lines represent 2-fold increase and decrease of that value. Dashed lines represent unstable steady states. c) Distribution of the critical signal strength beyond which the system jumps to the higher steady state for a population distributed as a lognormal with parameters $m=\log(1)$ and $s=0.25$. d) Several kinetic trajectories for the same population. Continuous black line represents the average cell of this population and dashed line the average of all trajectories. For all figures $k_0=0.4$, $k_1=0.01$, $k_2=k_3=1$, $k_4=0.06$, $K_{m1}=K_{m2}=0.05$.

In order to do so we rewired the previous module by adding a connection from the output to the signal input, hence creating a positive feedback loop. As expected, this induces a bistable system that above a critical signal exhibit a bifurcation in which only one of the two stable steady states persists, making the system switch abruptly to a higher state. However, due to the heterogeneity in the total amount of protein of the Koshland-Golbeter module, this critical signal strength is distributed across the population as can be seen in Figure 5 b). As in the previous example, for the same signal strength some cells will be above the threshold while others will not. Notice that the high steady state is poorly affected by the heterogeneity on the total concentration of the Koshland-Golbeter switch, its effect being felt primarily on the critical signal strength that induces the switch.

As such, this module allows the regulation of the number of cells that switch to a permanent state, making use of the intrinsic variability between cells.

Discussion

High throughput based systems biology holds the promise to disentangle the cell's circuitry in a systematic and straightforward way. This research program emerged with the advent of "omics" and established the top down approach as a major method in biological sciences. Starting from a global view of the system in question one hopes to infer the network of interactions that controls it, usually through computational approaches applied to bulk kinetic data upon perturbations to the system.

However, the realization that populations of cells are not homogeneous in terms of protein content poses additional challenges to the feasibility of this research program. Besides the typical hindrances in model identification, as many models display the same behavior, we are now faced with the extra challenge in which, due to the fact that each cell presents different dynamics, the bulk averages of these models are irreproducible by the model that generated them. It is then not warranted that any cell follows the dynamics of the mean of the population casting serious difficulties to parameter estimation and even to qualitative model inference.

Our approach suggests that this challenge can be overcome by quantifying the population heterogeneity, by making use of single cell techniques such as flow cytometry or confocal microscopy, and proceeding in several ways. One option would be to measure output distributions, given that the intervening molecules have already been identified, and identifying mechanisms that transform input distributions to that particular shape. As we've seen, different mechanisms have different signatures in terms of output distributions and, if the underlying heterogeneity is known, it might be possible to identify or at least to exclude some mechanisms. Another route would be to quantify the dynamics of the output distributions and from there infer the correct model and correctly estimate parameters. This route is more challenging, both experimentally and theoretically, as new estimation methods and experimental design methods need to be developed. However, a few steps have been given in this direction: high throughput flow cytometry has been used to quantify the variability in specific protein content in yeast (Ghaemmaghami et al., 2003) and single cell imaging techniques have been used to assess individual trajectories of certain processes (Cluzel et al., 2000; Sigal et al., 2006).

Much work has also focused in the control of heterogeneity (Thattai and van Oudenaarden, 2002; Raser and O'Shea, 2004) since it is usually thought to be detrimental to the cell. However, given its ubiquitous presence and the impossibility of control, perhaps it is more useful to consider a shift in point of view from the single average cell to the population of heterogeneous cells. In fact, a population of cells might benefit from having its constituents having distributed responses. In a multicellular organism, regulating the percentage of cells that respond to a particular stimulus might be advantageous in maintaining a pool of precursors and in single cell organisms, having diverse sensitivities to a signal in a population might allow the colony as a whole to be sensitive to a wider range of signals than any cell alone could or allow a population to endure unpredictable environmental fluctuations, as has been proposed before (Thattai and van Oudenaarden, 2004; Kussell and Leibler, 2005).

It is worth noticing that in this chapter we considered the heterogeneity between cells to be static, i.e. we assume the processes that we are modeling to be much faster than the ones that lead to appreciable changes to the total amount of the relevant molecules. This assumption certainly holds true in signal transduction pathways, where the processes involved are usually faster than transcription. However, this heterogeneity has been shown to be dynamic (Sigal et al., 2006), albeit extremely slow in many cases, and theoretically analyzed in several other works (Berg et al., 2000).

In this work we have concentrated on a few modules that are representative of the possible behaviors of distributed populations. We have seen that the nonlinear nature of many mechanisms within the cell, when combined with heterogeneity of protein abundances, can lead to remarkable heterogeneity of cell responses. In particular, when analyzing bulk responses at the population level, this heterogeneity can lead to remarkable deviations from the behavior of the mechanism that generated those dynamics, making it extremely difficult to correctly identify the underlying mechanism, let alone estimate parameters. As such, correct identification of mechanisms requires single cell measurements and at least correct identification of variability of the involved molecules within the population of interest.

References

- Alon, U. (2007). "Simplicity in biology." Nature **446**(7135): 497.
- Berg, O. G., J. Paulsson and M. Ehrenberg (2000). "Fluctuations and quality of control in biological cells: zero-order ultrasensitivity reinvestigated." Biophys J **79**(3): 1228-36.
- Cluzel, P., M. Surette and S. Leibler (2000). "An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells." Science **287**(5458): 1652-5.
- Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002). "Stochastic gene expression in a single cell." Science **297**(5584): 1183-6.
- Furusawa, C., T. Suzuki, et al. (2005). "Ubiquity of log-normal distributions in intracellular reaction dynamics." Biophysics **1**: 25-31.
- Ghaemmamghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.
- Goldbeter, A. and D. E. Koshland, Jr. (1981). "An amplified sensitivity arising from covalent modification in biological systems." Proc Natl Acad Sci U S A **78**(11): 6840-4.
- Han, J. D., N. Bertin, et al. (2004). "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." Nature **430**(6995): 88-93.
- Ingolia, N. T. and A. W. Murray (2007). "Positive-feedback loops as a flexible biological module." Curr Biol **17**(8): 668-77.
- Kholodenko, B. N., A. Kiyatkin, et al. (2002). "Untangling the wires: a strategy to trace functional interactions in signaling and gene networks." Proc Natl Acad Sci U S A **99**(20): 12841-6.
- Kitano, H. (2002). "Computational systems biology." Nature **420**(6912): 206-10.
- Kitano, H. (2002). "Systems biology: a brief overview." Science **295**(5560): 1662-4.
- Krishna, S., B. Banerjee, T. V. Ramakrishnan and G. V. Shivashankar (2005). "Stochastic simulations of the origins and implications of long-tailed distributions in gene expression." Proc Natl Acad Sci U S A **102**(13): 4771-6.
- Kussell, E. and S. Leibler (2005). "Phenotypic diversity, population growth, and information in fluctuating environments." Science **309**(5743): 2075-8.
- Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-4.
- Raser, J. M. and E. K. O'Shea (2005). "Noise in gene expression: origins, consequences, and control." Science **309**(5743): 2010-3.

- Sigal, A., R. Milo, et al. (2006). "Variability and memory of protein levels in human cells." Nature **444**(7119): 643-6.
- Thattai, M. and A. van Oudenaarden (2002). "Attenuation of noise in ultrasensitive signaling cascades." Biophys J **82**(6): 2943-50.
- Thattai, M. and A. van Oudenaarden (2004). "Stochastic gene expression in fluctuating environments." Genetics **167**(1): 523-30.
- Tyson, J. J., K. C. Chen and B. Novak (2003). "Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell." Curr Opin Cell Biol **15**(2): 221-31.
- Yang, L. and P. A. Iglesias (2006). "Positive feedback may cause the biphasic response observed in the chemoattractant-induced response of Dictyostelium cells." Syst Control Lett **55**(4): 329-337.

V

Fluctuation Structure and Adaptation Dynamics

The Role of Protein Fluctuation Structure on Adaptation Dynamics

Tiago Paixão* and Jorge Carneiro*

**Instituto Gulbenkian de Ciência*

Summary

Populations of isogenic cells display an appreciable variability in protein copy numbers. This variability has been attributed to the stochastic nature of intracellular events and has been shown to be dynamic, i.e., protein levels fluctuate within individual cells, with different rates. The diversity that results from this stochastic process opens the possibility of selection based on the amount of protein an individual cell contains. This leads to the question of whether the observed variability is beneficial for a population. Specifically, does higher variance of the steady state distribution of protein copy numbers confer an advantage? Do different steady state distributions of this variability have different outcomes in terms of somatic adaptation? If so, this raises the possibility of selection of genetic parameters that modulate the structure of these fluctuations.

Here, we assess the impact of different variances on the dynamics of somatic adaptation to a new environment. Moreover, we compare different steady state distributions in their capacity to facilitate such adaptation. We conclude that higher variances indeed provide an advantage, at least in certain conditions, and that different fluctuation structures (as assessed by the steady state distribution) have different abilities for somatic variation, raising the possibility for a modulation of the fluctuation by genetic parameters.

Introduction

Isogenic cell populations show a remarkable heterogeneity in the protein concentrations they express (Sigal et al., 2006). This heterogeneity has been explained as the result of stochastic events due to the low number of proteins involved in the mechanisms that lead to protein expression (McAdams and Arkin, 1997; Thattai and van Oudenaarden, 2001). Although it is commonly thought that stochasticity in protein expression is detrimental for cell function (Thattai and van Oudenaarden, 2002; Raser and O'Shea, 2004; Raser and O'Shea, 2005), recent works raised the possibility that the generation of phenotypic variants is in fact beneficial for bacterial populations, particularly when there is a need for fast adaptation to a randomly fluctuating environment (Thattai and van Oudenaarden, 2004; Kussell and Leibler, 2005). Previous works have focused mainly on exploring the consequences of stochasticity at the level of gene regulation, where it can induce transitions between discrete differentiation states of the population constituents, as in the case of the lysis/lysogeny decision in lambda phages (Arkin et al., 1998) or sugar utilization preferences in bacteria (Thattai and Shraiman, 2003). This process prompted further works (Kussell and Leibler, 2005) investigating the possibility that the switching rates between these states could be subject to selection.

The fact that stochastic events are responsible for protein expression entails that protein concentrations, such as receptor concentrations, fluctuate within each cell. In the specific case of receptors, these fluctuations alter the way each cell senses and responds to environmental signals without actually changing the genetic program or metabolic state of the cell, effectively creating a non-genetic meta-population. This raises the question of whether this spontaneous somatic variation can be exploited by cellular populations, for example, to increase the sensitivity of the population to a particular stimulus. More interestingly, it raises the question of how the dynamic characteristics of this heterogeneity affect the fitness of such population. On the other hand, this heterogeneity may carry undesirable consequences, such as uncertainty in determining local environment and non-optimal responses. This trade-off is clearly related to the response function the cells metabolism elicits upon receptor engagement.

Although parallels with mutation and selection dynamics are easily drawn from this scenario, a key difference needs to be outlined: somatic variation under this hypothesis is dependent on the current state of the cell, meaning that probabilities and magnitude of beneficial "mutations" is not constant; the farther off the

concentration of a protein within a cell is from the nominal value, the less likely it is to increase further its concentration.

Moreover, recent works focused on the specific shape of protein distributions observed in cellular populations and several explanations have been proposed for the lognormal distribution, which seems to be predominant (Furusawa et al., 2005; Krishna et al., 2005). Since different stationary distributions entail different kinetics for the fluctuation structure, we expect different stationary distributions to have different outcomes in competitions assays, hence providing basis for adaptation in terms of fluctuation structure.

Here we address the question of whether such fluctuations can play an adaptive role by simulating competitions between populations with different fluctuation structures of a receptor for proliferation signals in a typical competition assay.

Model description

We consider a population in exponential growth whose concentration of a receptor that is important in mediating proliferative signals is able to fluctuate according to the stochastic differential equation $\frac{dx}{dt} = f(x) + g(x)\xi(t)$, where $f(x)$ and $g(x)$ are arbitrary functions modeling the determinist dynamics of receptor expression and the stochastic perturbations to these, respectively and $\xi(t)$ is gaussian noise. At each instant the protein concentration within each cell is pulled towards its steady state value by the action of the deterministic “force” $f(x)$ but is subjected to random variation due to the stochastic “force” $g(x)$. In the case of an isogenic population, this gives rise to a population with distributed amounts of protein x . Each cell is then subjected to differential growth (meaning that cells with higher amounts of protein x proliferate more) with rate $k_g(x)$ that changes the shape of this distribution, as depicted in Figure 1.

Both processes can be formalized at the population level with the following approach. We make use of the Euler-Maruyama numerical scheme of solving SDEs in order to calculate the transition probability in time interval Δt :

$$P(X_{t+\Delta t} = x_j | X_t = x_i) = N(x_i + \Delta t f(x_i), \sqrt{\Delta t} g(x_i)) = M(x_j | x_i)$$

where $N(\mu', s')$ denotes a normal distribution with average $\mu'_i = x_i + \Delta t f(x_i)$ and standard deviation $s'_i = \sqrt{\Delta t} g(x_i)$. This gives the distribution at the next timestep of a homogeneous population whose receptor concentration was x_i . This population will then be subjected to exponential growth, modeled by multiplying this probability by

an exponential factor dependent on x_j and so the evolution of probability distribution of a population subjected to somatic variation and state dependent growth is given by:

$$p^{t+\Delta t}(x_j) = \int M(x_j|x_i) e^{k_G x_j} p(x_i) dx_i$$

the propagator $G(x_j|x_i) = M(x_j|x_i) e^{k_G x_j}$ can easily be shown to be a Gaussian distributed with $\mu_i = \mu'_i + k_G s'_i$ and $s_i = s'_i$, in the case where $k_g(x) = k_G x$. This propagator allows us to gain many insights into the dynamics and asymptotic behavior of populations that undergo this process. In particular, the average of this propagator denotes the average of a homogeneous population after these processes and can be used as measure of the displacement of the population relative to the population's average on the previous generation. Moreover, when this average is equal to the value of the population on the previous generation, it means that the population cannot adapt more and as attained variation-selection balance.

Results

Populations with higher variances have higher variation-selection equilibrium points

Because the steady state distribution of the fore mentioned SDE is described by:

$$\rho_{ss}(x) = \frac{1}{g^2(x)} e^{\frac{1}{2} \int \frac{f(x)}{g^2(x)} dx}$$

we can choose functions $f(x)$ and $g(x)$ such that the steady state distribution of the stochastic process, i.e. in the absence of differential proliferation hence free of selection, has a particular shape.

The simplest assumption for the fluctuation statistics of a molecule is Gaussian fluctuations. As such, we chose $f(x)$ and $g(x)$ to be $f(x) = k - dx$ and $g(x) = \sigma$. This entails a steady state distribution (when no selection effects are in place) that is Gaussian shaped with average k/d and variance $2\sigma^2/d$. Replacing these functions in the propagator's average we get $\mu^{t+\Delta t} = x_i + \Delta t(k - dx) + k_G(\sqrt{\Delta t}\sigma)^2$, which can be cast in the more appealing form:

$$\frac{\Delta x}{\Delta t} = k - dx + k_G \sigma^2 \quad (1)$$

This expression is the rate of change per generation of the average protein amount for a homogeneous population with protein amount x due to the variation and growth processes. From this expression we can analyze the rate of adaptation of any

population exhibiting Gaussian fluctuations, given a growth differential per generation (k_G).

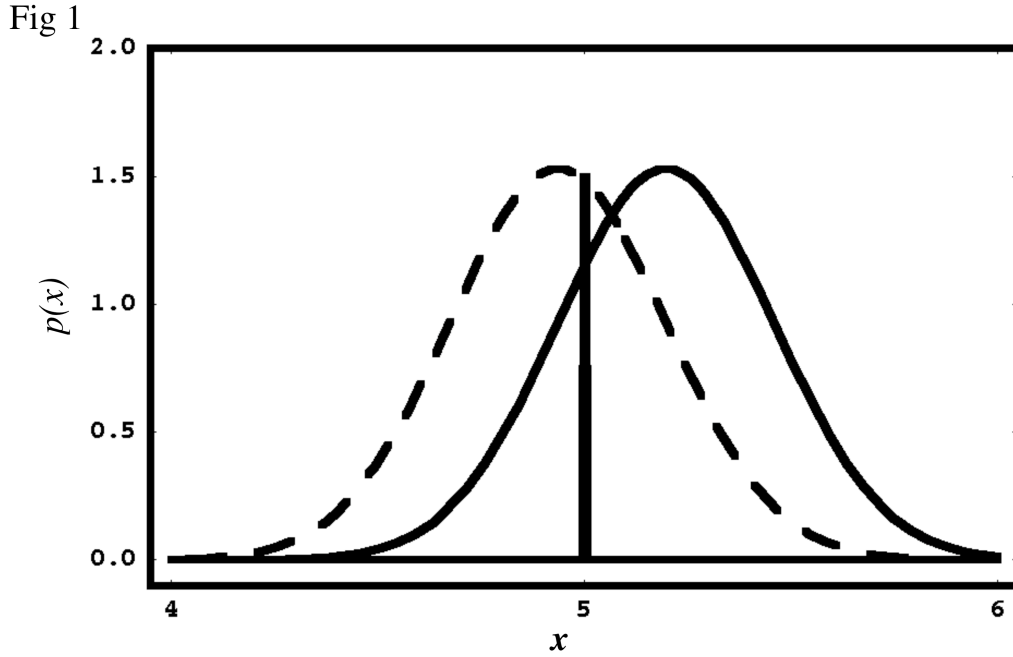


Figure 1: The variation growth process. A homogeneous population (vertical bar) suffers somatic variation which tends to “push” the population towards its nominal value (dashed line). Differential growth of the population constituents can then further reshape this resulting distribution increasing its mean.

Although larger variances have been shown to be beneficial for a population in a fluctuating environment and switching rates shown to be optimal when coupled to those of the fluctuating environment, the question of how populations with larger variances behave in a regular selecting environment needs to be addressed. In order to do so, we analyze how two populations exhibiting Gaussian fluctuations of a receptor for proliferation signals, albeit with different steady state variances behave in such an environment. We consider a nominal population which fluctuates with parameter k_1, d_1 and σ_1 , and a second population whose steady state variance is half of the nominal population with parameters $k'=k, d'=d$ and $\sigma_2^2 = \sigma_1^2/2$. As can easily be checked, the variance of this second population is half of the nominal one although its average is the same. As can be seen on Figure 2a) this latter population attains variation-growth equilibrium at a point much lower than the nominal population. In fact, by analyzing equation (1) in both cases (Figure 2c) we see that the population with higher variance has a higher stabilizing point $\left(\frac{k}{d} + \frac{k_G \sigma^2}{d} > \frac{k}{d} + \frac{k_G \sigma^2}{2d} \right)$ and the same

slope than the lower variance population. As such, this low variance population adapts slower than the high variance one and always stabilizes at a lower point, meaning that it would be out competed.

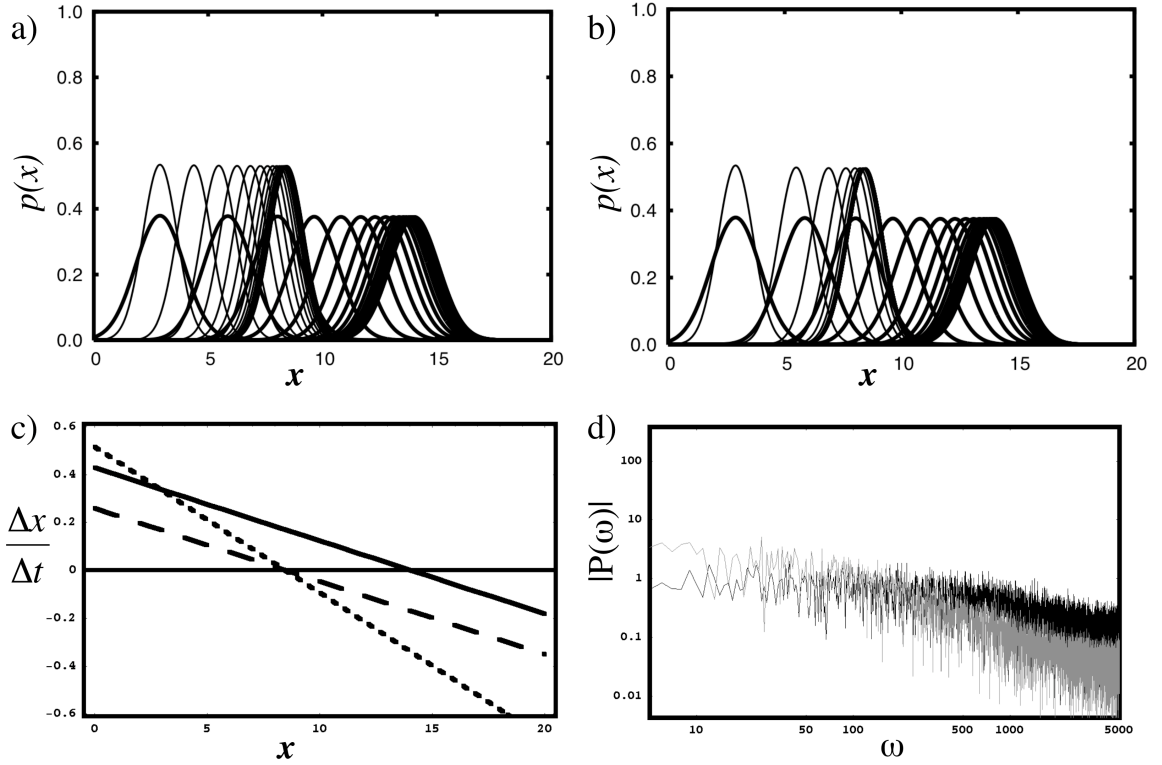


Figure 2: Different variances have an impact on the dynamics of adaptation. Competitions between a nominal population and one that has half its variance by having kinetic parameters a) $\sigma'=\sigma/\sqrt{2}$ or b) $k'=2k$ and $d'=2d$, invariably leads to the prevalence of the higher population. c) adaptation rates for the nominal population (continuous line) and for the less broad populations as in a) (long dashes) and b) (short dashes). d) Fourier spectrums for a timeseries of a cell of population a) (grey) and b) (black).

However, as can easily be seen, the previous lower variance population parameters are not the only solution for a population with such a variance. Another solution would be the set of parameters $k_2=2k_1$, $d_2=2d_1$ and $\sigma_2=\sigma_1$. This population has the same stabilizing point than the previous population but has a higher absolute slope than its high variance counterpart. This means that initially the lower variance population has a higher rate of adaptation, although losing its advantage as it stabilizes at a lower point than the higher variance population. As we are considering an infinite population this means that the high variance population always wins the competition. Although, the existence of a region where the low variance population has an adaptive advantage means that in a finite population the high variance

population might be led to extinction if its numbers do not support the decline that it stands until it is able to escape that region.

The fact that, although both distributions have the same steady state distribution, they have different dynamic characteristics, prompted us to explore more systematically how the dynamics of fluctuation influence the rate of adaptation. We can model a population whose fluctuation structure is slower, although otherwise equal, by making the following substitution to the SDE functions of the reference population: $f_s(x) = \alpha f(x)$ and $g_s^2(x) = \alpha g^2(x)$. As can easily be seen, the propagator's intercept is the same for both populations but the slope is different, meaning that a slower population will always adapt slower than a faster one, although having the potential to evolve to the same point.

Different steady state distributions entail different adaptation dynamics

The fact that the two low variance populations of the previous section show distinct adaptation kinetics shows that the steady state distribution is not the only determinant of dynamic characteristics. In fact, the two lower variance population show distinct dynamic characteristics as can be seen by their respective Fourier spectrums on Figure 2d). This prompted us to explore the role of different steady state distributions on these competitions.

In order to do so, we chose functions $f(x)$ and $g(x)$ such that we can obtain populations with gamma and lognormal distributed receptors. In order to obtain gamma distributed populations we can choose $f(x) = f_g(x) = k_g - d_g x$ and $g(x) = g_g(x) = \sqrt{\sigma_g} x$. This entails a steady state protein distribution which is gamma distributed with parameters $\alpha = \frac{k_g}{2\sigma_g}$ and $\theta = \frac{2\sigma_g}{d_g}$ and with average $\mu_g = \alpha\theta = \frac{k_g}{d_g}$ and variance $\alpha\theta^2 = \frac{2\sigma_g k_g}{d_g^2}$.

The rate of adaptation of the propagator for this distribution is then:

$$\frac{\Delta x}{\Delta t} = k_g - (d_g - k_g \sigma_g) x$$

This equation shows that this rate has the same shape as the Gaussian distribution, a straight line, with the difference now the growth differential has influence on the slope of this line. It also shows that the possibility exists that, for high enough values, could turn the slope to a positive one, meaning that under these conditions the population would undergo a runaway process, adapting with increasing rate. Still, for

each set of parameters, the rate of adaptation for a gamma distribution is monotonic and a population with slower fluctuations always adapt slower to the equilibrium point, as can be seen on Figure 3.

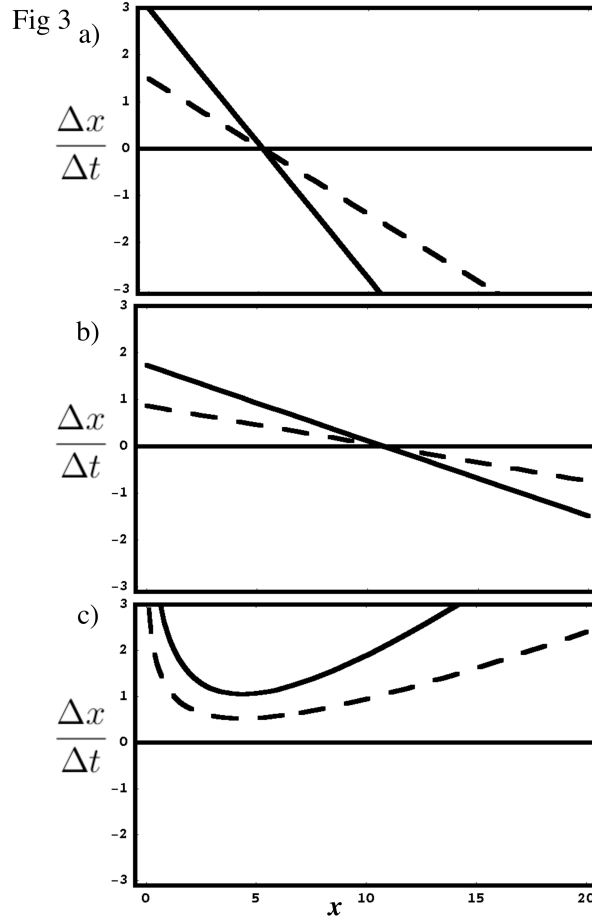


Figure 3: Adaptation rates for populations with different steady state distributions but with the same average ($\mu=\sqrt{3}$), variance ($3/4$) and dispersion rate around the mean ($1/2$). a) gaussian distribution. b) gamma distribution and c) lognormal distribution. Dashed lines are the same distributions with half the fluctuation rate.

In order to compare different distributions as fairly as possible, we require both their averages and their variances to be the same and additionally, their rate of dispersion around the average: $g_g(x_m)$. Using these constraints, we can cast the parameters of the gamma dynamics in terms of the Gaussian parameters:

$$k_g = k ; d_g = d ; \sigma_g = \sigma^2 \frac{d}{k}$$

Using these parameters, it is easy to see that the adaptation rate for the gamma dynamics has a lower intercept with the yy axis (k vs $k+k_g\sigma^2$) and a lower slope ($d-k_g\sigma d/k$ vs d) meaning that its steady state value can be higher or lower than the Gaussian one, depending on the specific values of these parameters.

A lognormal shaped steady state distribution is obtained by choosing $f_l(x) = k_l - d_l \log(x)$ and $g_l(x) = \sqrt{\sigma_l} x$, which entails a lognormal distribution with average $\mu_l = e^{M + \frac{S^2}{2}}$ and variance $e^{2M + S^2} (e^{S^2} - 1)$ and parameters $M = k_l/d_l$ and $S^2 = 2\sigma_l/d_l$. Substituting these functions in eq. 1 we can compute the rate of adaptation of the propagator for a lognormal distribution:

$$\frac{\Delta x}{\Delta t} = k_l - d_l \log(x) + k_G \sigma_l x$$

In contrast with the Gaussian and the gamma distributions, the rate of adaptation for a process that follows lognormal statistics is not a straight line, due to the presence of the logarithmic term. In fact, these dynamics show biphasic behaviour, exhibiting a decreasing rate of adaptation while the logarithmic term dominates and an increasing rate of adaptation as the linear term predominates. This confers a complex behaviour to the adaptation of populations that display lognormal dynamics. These can range from non-stabilizing dynamics, where the population undergoes a runaway process where the protein levels are always increasing, to a situation where the rate of adaptation has 2 steady states, one of which (the lower) is unstable.

As such these different distributions have different dynamic characteristics of adaptation. For the specific parameter set chosen in Figure 3, we can see that the lognormal distribution has an advantage since it doesn't have an equilibrium point. However, decreasing the growth differential makes the lognormal dynamics acquire an equilibrium point which, although initially is higher than the one of the gamma dynamics, eventually becomes lower than the gamma equilibrium, as k_G is further lowered. Of course different parameter sets have different behaviours, and the nonlinearity of the expression of the lognormal rate of adaptation prevents a more analytical approach.

Discussion

In this paper, we analyze the effect of different fluctuations statistics of a protein that has a direct effect on the fitness of a population on the growth characteristics of a population.

Previous works focused on the selective advantage of several different metabolic states in the context of a fluctuating environment. In these works, the switching rates between these states is disconnected from any particular mechanism and it was assumed that the environment entailed an optimum cellular state, whereby the

switching rates were a way for the population to “predict” the fluctuations of the environment. In contrast, we show here that under a regular selecting environment, the fluctuation rates and steady state distributions of a specific molecule have an impact on the fitness of the population. We found that the structure of these fluctuations can play a determinant role in the dynamics of adaptation of a population. In particular we show that higher variances (for the same average value) always show faster kinetics and, consequently, higher points of somatic variation/growth balance. Also, we have show that faster fluctuation rates always display higher adaptation rates to the same equilibrium point. This suggests that there is selective pressure for larger variances in competing populations.

The fact that different distributions entail different adaptation dynamics raises the question of which steady state distribution is more beneficial for the population, particularly since the lognormal distribution is commonly observed in cell populations. A multitude of factors seem to affect this question. First, the magnitude of the maximum adaptation attainable (if it is attainable) is linked to the differential growth that the population is subjected. Second, the adaptation rate is linked to the population size, since smaller adaptation rates even with higher adaptation steady states, could be out competed due to the finite nature of the elements of the population. As such, the question of what is the optimal distribution for a population is a question about the magnitude of selective pressures (differential growth) that the population is subjected to, its population size and what populations it is competing with.

The Gaussian and gamma distributions show almost the same behaviour for small k_G 's, with their adaptation steady states departing for higher differential growth rates. However, for higher k_G 's the gamma dynamics are able to disrupt this equilibrium, making the population increase their receptor levels without bounds. The specific case of the lognormal is interesting as it shows a very sensitive response to the value of k_G and two steady states. It is possible that, due to the long tail of this distribution, some cells would be able to jump to the unstable region beyond the higher steady state, providing a way to regulate the number of cells that undergo further adaptation.

This question of the optimal steady state distribution is also intimately linked to the question of the rate of fluctuations and its origin. Can these parameters be tuned by genetic factors, providing a basis for selection at that level? It has been shown that, in a simple model of protein expression (Ozbudak et al., 2002), the average and variance (and, consequently their dynamic characteristics) can be regulated by tuning of the transcription and translation rates. However, these parameters are not

specific for individual proteins, rendering their usage to tune individual protein fluctuation rates unlikely. However, in order to achieve a fully functional protein many posttranslational modifications are usually required. It is possible that these could be combinatorially tuned in order to give specific fluctuation rates. Moreover, for a receptor, in particular, turnover rates depend on interaction with specific enzymes, providing a basis for selectively tuning of these fluctuation rates. Further characterization and elucidation of the specific mechanism that generates this dynamic heterogeneity is needed in order to answer these questions.

In this work we assumed, for mathematical simplicity, that the fitness function was always increasing, meaning that the levels of receptor and its effect on the growth rate do not saturate. However, taking into account saturation of the effect of the receptor on the adaptation rate means that above a certain threshold amount of receptor the population defaults to the rate of change of its mean given by the variation process alone. This makes further adaptation impossible since without differential growth no reshaping of the distribution can occur. As such, a population with a saturation threshold below the variation-growth equilibrium would stabilize at some point between the two. It is also interesting to note that if populations are “pushed” beyond their equilibrium point (for instance, by being transiently subjected to a higher differential growth) the return to the equilibrium point is slower for slower populations, granting them an advantage.

References

- Arkin, A., J. Ross and H. H. McAdams (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells." Genetics **149**(4): 1633-48.
- Furusawa, C., T. Suzuki, et al. (2005). "Ubiquity of Log-normal Distributions in Intracellular Reaction Dynamics." Biophysics **1**: 25.
- Krishna, S., B. Banerjee, T. V. Ramakrishnan and G. V. Shivashankar (2005). "Stochastic simulations of the origins and implications of long-tailed distributions in gene expression." Proc Natl Acad Sci U S A **102**(13): 4771-6.
- Kussell, E. and S. Leibler (2005). "Phenotypic diversity, population growth, and information in fluctuating environments." Science **309**(5743): 2075-8.
- McAdams, H. H. and A. Arkin (1997). "Stochastic mechanisms in gene expression." Proc Natl Acad Sci U S A **94**(3): 814-9.

- Ozbudak, E. M., M. Thattai, et al. (2002). "Regulation of noise in the expression of a single gene." Nat Genet **31**(1): 69-73.
- Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-4.
- Raser, J. M. and E. K. O'Shea (2005). "Noise in gene expression: origins, consequences, and control." Science **309**(5743): 2010-3.
- Sigal, A., R. Milo, et al. (2006). "Variability and memory of protein levels in human cells." Nature **444**(7119): 643-6.
- Thattai, M. and B. I. Shraiman (2003). "Metabolic switching in the sugar phosphotransferase system of Escherichia coli." Biophys J **85**(2): 744-54.
- Thattai, M. and A. van Oudenaarden (2001). "Intrinsic noise in gene regulatory networks." Proc Natl Acad Sci U S A **98**(15): 8614-9.
- Thattai, M. and A. van Oudenaarden (2002). "Attenuation of noise in ultrasensitive signaling cascades." Biophys J **82**(6): 2943-50.
- Thattai, M. and A. van Oudenaarden (2004). "Stochastic gene expression in fluctuating environments." Genetics **167**(1): 523-30.

VI

Stochastic Cell-Cell Interactions as a Generator of Somatic Variation Within Lymphocytes

This chapter is partly based on the paper entitled “Immunological Self-Tolerance: Lessons from Mathematical Modeling” that was published on Journal of Computational and Applied Mathematics, (2005) 184: 77–100. Although the main message of this article is directed towards the understanding of the mechanism governing self-tolerance in the immune system, it presents several characteristics that make it relevant for the present thesis.

In order to model a specific hypothesis of the regulation of homeostasis of T lymphocytes, we formalize a model in which T cells interact with antigen presenting cells (APCs) in random contacts. In this model, the dynamics of contacts between T cells and APCs influence the latter’s intracellular state, which in turn influences the T cells population dynamics, by regulating their proliferation. Thus, it provides an example on how extracellular events can lead to heterogeneity in a population and how this propagates to the population dynamics in a nontrivial way. It shows how heterogeneity in a population can be exploited by a multicellular organism for the regulation and homeostasis of its constituent populations. Furthermore, this model provides a framework that allows the modeling of several organizational levels simultaneously. This is crucial for the further development of our understanding of the impact of intracellular events on population and community dynamics.

Immunological Self-Tolerance: Lessons from Mathematical Modeling

Jorge Carneiro¹, Tiago Paixão¹, Dejan Milutinovic², João Sousa¹, Kalet Leon^{1,3},
Rui Gardner¹, and Jose Faro¹

Instituto Gulbenkian de Ciência, Oeiras

Instituto Superior Técnico, Lisboa

Centro de Inmunología Molecular, Habana

Summary

One of the fundamental properties of the immune system is its capacity to prevent autoimmune disease by control of autoreactive lymphocytes. The mechanism underlying this process, known as self-tolerance, is hitherto unresolved. This article reviews mathematical modeling of self-tolerance, addressing a specific hypothesis. This hypothesis posits that self-tolerance is mediated by tuning of activation thresholds, which makes autoreactive T lymphocytes ineffective or “anergic”. A Model representing the population dynamics of autoreactive T lymphocytes according to this hypothesis was derived. We identified how cell density affects tolerance, and predicted the corresponding phase spaces and bifurcations. We show that the induction of anergy, as modeled, has a density dependence that is only partially compatible with adoptive transfers of tolerance, and that the models of tolerance mediated by specific regulatory T cells are closer to observations.

Introduction

Mathematical modeling of the immune system often concentrates on the immune responses to pathogens. This article deals with another fundamental process in the immune system: the maintenance of self-tolerance, i.e. the prevention of harmful immune responses against body components. The biological significance of this process becomes very patent upon its failure during pathological conditions known as autoimmune diseases.

The risk of autoimmunity cannot be dissociated from the capacity of the immune system to cope with diverse and fast evolving pathogens (Langman and Cohn, 1987). The latter is achieved by setting up a vast and diverse repertoire of antigen receptors expressed by lymphocytes, which as a whole is capable of recognizing any possible antigen. Most lymphocytes have a unique antigen receptor (immunoglobulin in B-cells and TCR in T cells) that is encoded by a gene that results from somatic mutation and random assortment of gene segments in lymphocyte precursors. The randomness in the generation of antigen receptors makes it unavoidable that lymphocytes with receptors recognizing body antigens are also made. These autoreactive lymphocytes can potentially cause autoimmune diseases. The question is how is this avoided in healthy individuals?

According to Burnet's original clonal selection theory (Burnet, 1957) autoreactive lymphocytes would be deleted once and for all during embryonic development. The fact that the generation of lymphocytes is a life long process in mammals invalidated this possibility. Following an early suggestion by Lederberg (Lederberg, 1959) deletion of potentially self-destructive lymphocytes was reformulated as an aspect of lymphopoiesis. Accordingly, lymphocytes that express an autoreactive receptor are deleted at an immature stage of their development (Kisielow et al., 1988; Acha-Orbea and MacDonald, 1995), before they can trigger destructive immune responses.

But deletion alone cannot explain self-tolerance. The major shortcoming of deletion models of self-tolerance is the well-documented presence of mature autoreactive B and T lymphocytes in normal healthy animals (Pereira et al., 1986). Many different experiments have demonstrated that these autoreactive T cells can cause disease. In this paper we will focus on a particular type of experiments first reported by Sakaguchi and coworkers (Sakaguchi et al., 1995; Sakaguchi et al., 2001), which allows assessing self-tolerance from the perspective of the population

dynamics of circulating lymphocytes. CD4⁺ T cells were isolated from healthy animals and subsets of this population were transferred into syngeneic recipient animals, which were devoid of T cells. Transfer of CD4⁺CD25⁻ T cells caused in the recipients an autoimmune syndrome characterized by multiple organ-specific autoimmune diseases (illustrated in fig. 1). These results indicate that in the healthy individuals there are significant numbers of autoreactive cells that could potentially mount deleterious immune responses to self.

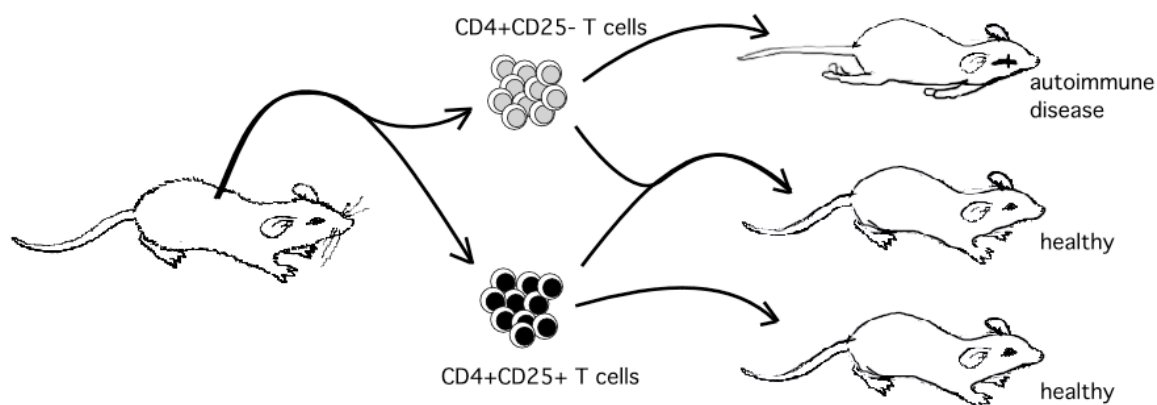


Figure 1 – Illustration of the experiments of Sakaguchi et al. demonstrating the existence of autoreactive T cells in healthy individuals. Purified CD4⁺CD25⁻ T cells, but not CD4⁺CD25⁺ T cells, from health animals will cause autoimmune diseases in recipient animals. An interaction between the two subsets of CD4 cells prevents disease development in the same recipients.

How are those autoreactive T cells, circulating in healthy individuals, prevented from mounting harmful immune responses against body tissues? There are several hypotheses in the literature (see the special issue of *Seminars Immunology* (vol 12 issue 3) for a rather comprehensive overview). One hypothesis posits that autoreactive T cells are prevented from mounting immune responses because they are controlled by specific regulatory T cells. In the above-mentioned Sakaguchi et al. experiment (fig. 1), those animals devoid of T cells receiving the same number of CD4⁺CD25⁺ T cells or receiving equal numbers of CD4⁺CD25⁻ and CD4⁺CD25⁺ T cells did not develop autoimmune diseases. Prevention of autoimmunity in the recipients by transfer of CD25⁺ T cells suggests the existence of regulatory T cells within the CD25⁺ subset, which exert a direct suppressive interaction on CD25⁻ T cells. Although this interpretation has been favored by immunologists, recent evidence indicates that competition and density-dependent inhibition of cell expansion in recipients may be sufficient to explain the inhibitory effects of CD4⁺CD25⁺ cells on CD4⁺CD25⁻ cells (Barthlott et al., 2003), and thus

postulating direct suppressive effects could be superfluous. Another hypothesis for the prevention of harmful immune responses by autoreactive T cells is that these cells become unresponsive to self-antigens by modification of their cell-signaling machinery; immunologists refer to this unresponsive state of the individual cell as anergy. Among the possible explanations for self-specific anergy induction, perhaps the simplest is the hypothesis that lymphocytes tune up their activation thresholds in response to recurrent stimuli (Grossman and Paul, 1992; Grossman and Paul, 2000; Grossman and Paul, 2001). According to this tunable activation threshold (TAT) hypothesis, autoreactive lymphocytes that are frequently stimulated by particular self-antigens adapt to the recent time-average of such stimulation so that they fail to be activated by these antigens.

In this article we present a simple model for the dynamics of autoreactive T cells when endowed with a mechanism that allows them to adapt their activation threshold based on the frequency of presentation of antigen. Basically, we ask here whether and to what extent the underlying mechanisms of tolerance induction by tuning of activation thresholds are compatible with the basic aspects of the Sakaguchi phenomenon (fig. 1). The next section addresses the hypothesis that recurrent stimulation by self-antigens and tuning of the activation threshold regulates the dynamics of autoreactive T lymphocytes. We show that the induction of anergy/tolerance by tuning, as implemented in our model, has a density dependence that is only partially compatible with the Sakaguchi phenomenon.

Modeling tolerance by tuning of activation thresholds of individual T lymphocytes

The tunable activation threshold (TAT) hypothesis by Grossman and Paul (Grossman and Paul, 1992; Grossman and Paul, 2000; Grossman and Paul, 2001) proposes that every interaction between the TCR and its ligand on APCs results in an intracellular competition between "excitation" and "de-excitation" signaling pathways that causes the T cell to adapt to the stimulation by increasing or decreasing its threshold for activation. Recently, this hypothesis has been discussed in the context of peripheral tolerance and homeostasis (Nicholson et al., 2000; Smith et al., 2001; Tanchot et al., 2001; Wong et al., 2001; Singh and Schwartz, 2003). The TAT hypothesis is certainly inspired on tuning and adaptation of neural synapses.

However, the set up of the immune system poses specific problems. In neural tissue intermittent signaling does not depend on de novo formation of synapses, which are sessile. In contrast, the immunological synapses are intermittent and their formation depends on the relative densities of T cells and APCs bearing agonist antigens. Synapse formation and signaling is therefore coupled to population dynamics and, as we will see below, poses novel modeling problems.

A Minimal Model

Activation, proliferation and survival of T lymphocytes require recurrent interactions of their TCRs with their ligands, the MHC-peptide complexes, at the membrane of antigen presenting cells (APCs) (Witherden et al., 2000; Polic et al., 2001). This APC-dependent population dynamics is captured in the reactional diagram proposed by De Boer and Perelson (De Boer and Perelson, 1994) (fig. 2A). Assuming that the densities of the conjugated T cells are in quasi-steady state (De Boer and Perelson, 1994; Sousa, 2003), this diagram is translated into the following differential equation:

$$\frac{dT}{dt} = d\alpha C - \delta T \quad (1)$$

where T is the T cell density, d is the conjugate dissociation rate, δ is the per cell death rate, and C is the quasi-steady state conjugate density:

$$C \equiv \frac{c(T + A) + d - \sqrt{-4 \cdot A \cdot T \cdot c^2 + (c(T + A) + d)^2}}{2c} \quad (2)$$

According to the TAT hypothesis continuous signaling by the TCR would lead to dynamic adaptation of the signaling transduction machinery. To incorporate this in the model above we must define the probability α of productive conjugation as a function of the signaling machinery status of the conjugated lymphocyte. Sousa et al. (Sousa, 2003) proposed that T cell activation is controlled by a “futile” cycle downstream of TCR signaling, involving a kinase and a phosphatase that operate on an adapter molecule (Clements et al., 1999; Germain and Stefanova, 1999; Stefanova et al., 2003) (fig. 2). In a setup inspired by (Koshland et al., 1982), it is assumed that the phosphorylation state of the adapter is hypersensitive to the relative activities of the two enzymes and behaves as a molecular switch. All the adapter molecules are phosphorylated if the kinase activity is higher than that of the phosphatase; otherwise, the adapter is fully dephosphorylated. During lymphocyte

conjugation with an APC, TCR stimuli result in a faster increase in kinase and phosphatase activities. The lymphocyte will be activated and enter cell cycle if after conjugating with the APC the kinase activity supersedes that of the phosphatase, and it will remain quiescent otherwise. The dynamics of this signaling machinery was represented by two differential equations:

$$\frac{dK}{dt} = r_K(K_0(1 + \sigma) - K) \quad (3)$$

$$\frac{dP}{dt} = r_P(P_0(1 + \sigma) - P) \quad (4)$$

where K is the kinase activity, P is the phosphatase activity, r_K is the turnover rate of the kinase, r_P is the turnover rate of the phosphatase, K_0 is the basal rate of production of kinase activity, P_0 is the basal rate of production of phosphatase activity, and σ is the magnitude of the stimulus to the kinase and phosphatase production rates. This signaling cascade shows the adaptive properties posited by Grossman et al. (Grossman and Paul, 1992; Grossman and Paul, 2000; Grossman and Paul, 2001) provided that the turnover rate of the kinase is higher than that of the phosphatase ($r_K > r_P$), and that, for any stimuli, the steady state activity of the phosphatase is higher than that of the kinase ($P_0 > K_0$). Under these conditions, the adapter can be transiently switched on, but it will be switched-off eventually if the stimulus persists.

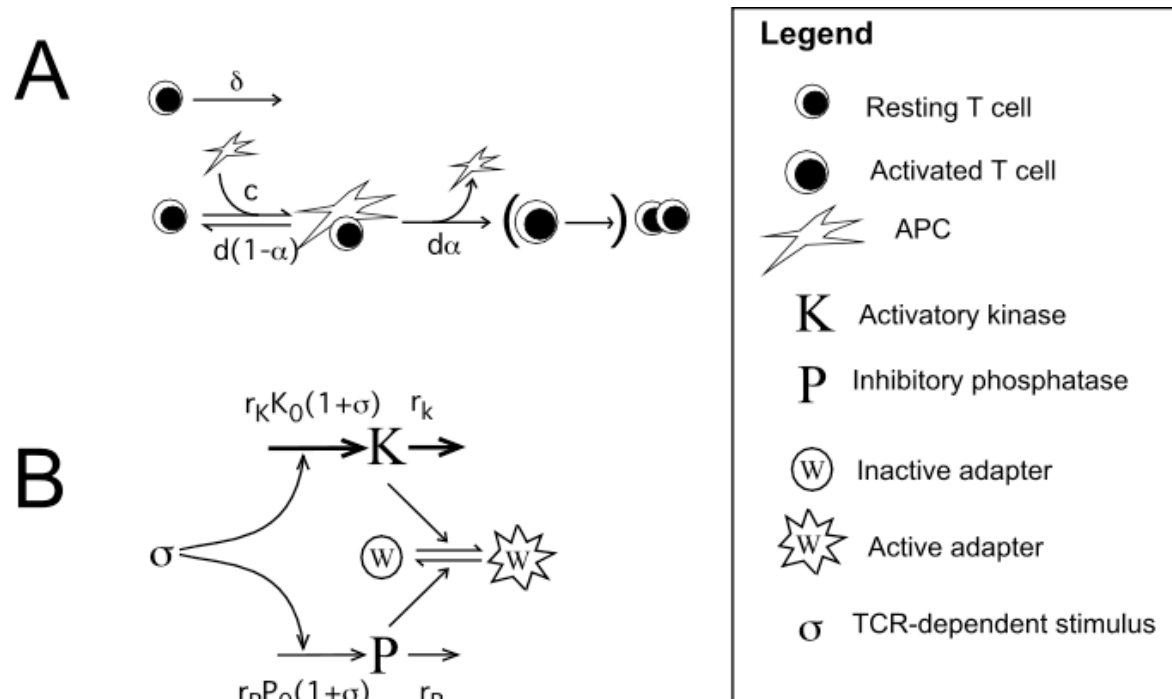


Figure 2 – Illustration of a model of the population dynamics of lymphocytes with tunable

activation thresholds indicating the cellular processes (A) and the molecular processes of the cell machinery (B).

This simple TAT model was proposed and analysed in Sousa et al. (Sousa, 2003), based mainly on Monte-Carlo stochastic simulations of individual cells. In this article we present a further simplification, which is amenable to analytic treatment and retains the main properties of the original model. This simplification involves two additional approximations. First, we assume that turnover of the kinase activity is very fast as compared to the conjugate dissociation rate ($r_K \gg d$), and as compared to the turnover rate of the phosphatase activity ($r_K \gg r_P$). Under these conditions, the kinase activity is in quasi-steady state, and it can be approximated by either $K = K_0(1+\sigma)$ or $K = K_0$, respectively, when the T lymphocyte is conjugated to an APC resulting in a stimulus σ or when the lymphocyte is free. The second approximation consists in assuming that for any given density of T cells and APCs, the fast conjugation and deconjugation processes are practically in equilibrium. This implies that the probability density functions (pdf) of the phosphatase activity in conjugated and in free T cell populations are stationary.

Milutinovic et al. (Milutinovic et al., 2004) used stochastic hybrid-automaton theory to describe the pdfs of cell-associated molecules in cells that cycle between APC-conjugation and APC-free states. The dynamics of phosphatase pdf in conjugated and free T cells, respectively ρ_C and ρ_F , are described by the following set of first order partial differential equations:

$$\frac{\partial \rho_C}{\partial t} + \frac{\partial}{\partial P}(P_C \rho_C) = -d\rho_C + c_E \rho_F \quad (5)$$

$$\frac{\partial \rho_F}{\partial t} + \frac{\partial}{\partial P}(P_F \rho_F) = d\rho_C - c_E \rho_F \quad (6)$$

where P_C and P_F are the phosphatase dynamics in the conjugated and free regimes, respectively $r_P(P_0(1+\sigma)-P)$ and $r_P(P_0-P)$, and c_E and d are the transition rates from the free to the conjugated state ($c_E=c(A-C)$) and from the conjugated to the free state, respectively.

Assuming that the conjugation and deconjugation processes are in quasi-steady state, we expect the time derivatives to vanish. Under these conditions, we obtain the following equation:

$$\frac{\partial}{\partial P}(P_C \rho_C + P_F \rho_F) = 0 \quad (7)$$

which, as demonstrated by Milutinovic et al. (Milutinovic et al., 2004), can be used to reduce the system to the following differential equation:

$$\frac{\partial}{\partial P}(P_C \rho_C) = -\left(d + c_E \frac{P_C}{P_F}\right) \rho_C$$

(8)

The solution of this equation is:

$$\rho_C = \begin{cases} N \left[P_0(1 + \sigma) - P \right]^{\frac{d}{r_P} - 1} \left[P_0 - P \right]^{\frac{c_E}{r_P}}, & P_0 \leq P \leq P_0(1 + \sigma) \\ 0, & \text{else} \end{cases} \quad (9)$$

where N is a normalisation constant.

The fraction α of T cells that is activated and divides when conjugation ceases (i.e. the fraction of T cells that at the instant of releasing from the APC have $K > P$) is then:

$$\alpha = \int_{P_0}^{K_0(1+\sigma)} \rho_C dP \quad (10)$$

Substituting this definition in eqn. 1 we fully define the population dynamics of the T cells with TAT.

Results

The (K, P) -signaling machinery shows an activation threshold that is tunable. It is easy to note that if the value of the phosphatase activity at the beginning of conjugation is $P \geq K_0(1 + \sigma)$ then this is sufficient (although not necessary) to prevent activation of the T cell when ligand concentration is $\leq \sigma$. The threshold is modulated by the history of stimuli to the T cell, which determines the value of the phosphatase activity P at any given time. Therefore, from the point of view of the population biology of T cells, in this model the activation threshold is dependent on the frequency of interactions of T cells with the APCs, i.e. the frequency of the stimuli to the individual T cells (Sousa, 2003) (fig. 3A). If T cells were always free, the probability density function of the phosphatase activity would correspond to a Delta Dirac at P_0 ; while if all the lymphocytes were permanently conjugated to APCs delivering the same stimulus σ then the pdf would be a Delta Dirac at $P_0(1 + \sigma)$ (fig. 3B). Since T cells are cycling between conjugation and free periods, the stationary pdf of P in the population takes values in the interval $[P_0, P_0(1 + \sigma)]$ (fig.3B).

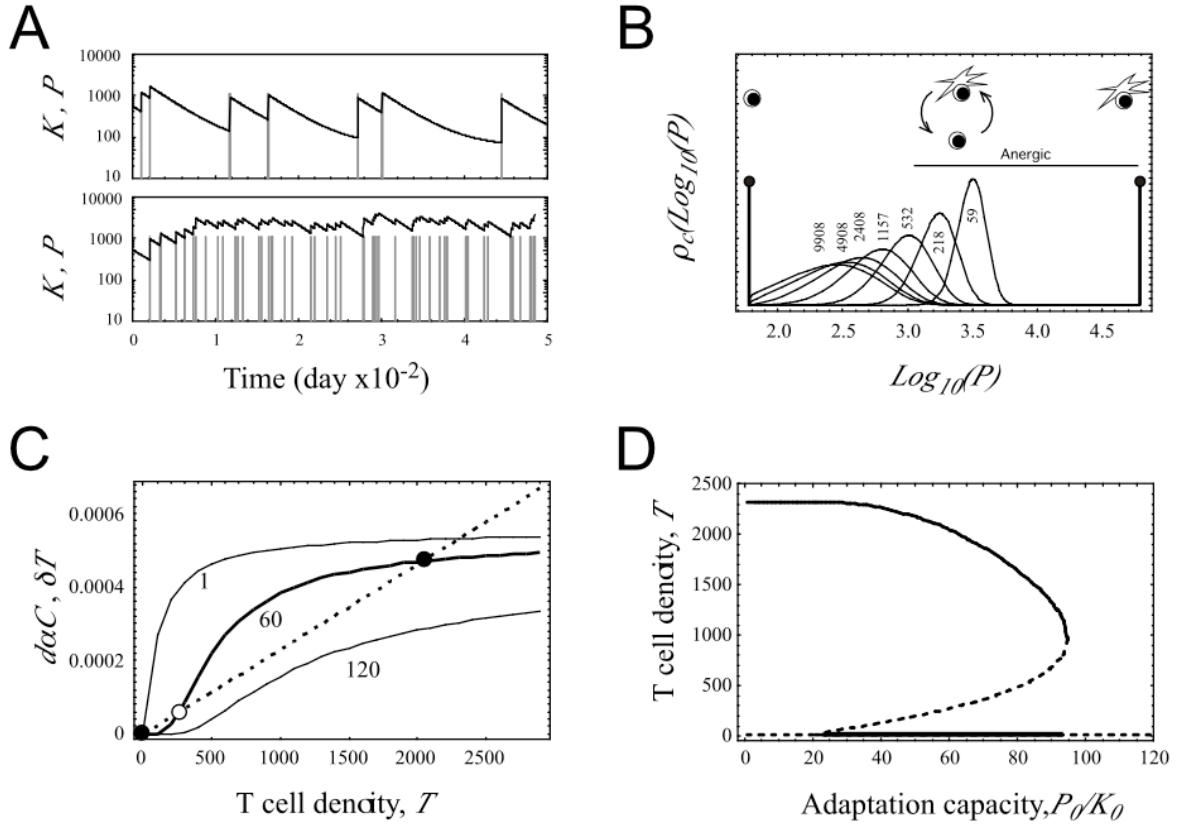


Figure 3 – Analysis of the model of the population dynamics of T cells with tunable activation thresholds. A- Kinetics of the phosphatase (black line) and the kinase (gray line) which are downstream of TCR stimulus in an individual T cell. The probability that the phosphatase activity P supersedes the kinase activity K at the instant of deconjugation increases with the frequency of encounters. (Top: $c(A-C)=0.20$; bottom: $c(A-C)=0.048$). B- Stationary probability density function of the phosphatase activity P in populations of T cells at the indicated densities ($P_0 = 60$; $\sigma = 1000$). C- Phase diagram of the model indicating the death rate (dashed line) and growth rates (solid lines) for the indicated values of the control parameter P_0/K_0 ; the dots indicate the stable (black) and unstable (white) steady states for the reference value $P_0/K_0=60$. D- Bifurcation diagram obtained by varying the control parameter P_0/K_0 that determines the adaptation capacity of the signaling machinery; Stable and unstable steady states are indicated by solid and dashed lines respectively. (Reference parameters: $r_P=0.027 \text{ day}^{-1}$; $K_0=1 \text{ au}$; $P_0=60 \text{ au}$; $\sigma=1000$; $A=8 \text{ cells}$; $d=6 \text{ day}^{-1}$; $c=0.06 \text{ cell}^{-1}\text{day}^{-1}$; $\delta=0.02 \text{ day}^{-1}$; $\text{au}=\text{relative activity units}$).

The frequency of APC interactions per T cell decreases as T cell density increases due to competition (eqn 2). This implies that, as T cell density increases, the median of the pdf of phosphatase activity in conjugated T cells ($\rho_c(P)$) becomes closer to the value P_0 ; reciprocally, as the T cell density decreases, the median of the pdf approaches $P_0(1 + \sigma)$ (fig.3B). This means that the fraction of cells α undergoing productive conjugation to activation and cell cycle increases with T cell density. This defines a positive feedback loop such that increases (decreases) in T

cell density result in higher (lower) average values of α , which lead to further increases (decreases) in T cell density. This positive feedback loop resulting from the present implementation of tunable activation thresholds is the opposite of a density-dependent feedback population control. In our model this loop interacts with the negative feedback loop defined by the effect of competition on the density of conjugates. For this reason, the model has two possible stable steady states: one in which lymphocyte population is extinct and one in which it is limited by APC availability, and predominantly made of non-anergic lymphocytes. The bifurcation diagram of the steady state population size as a function of the ratio P_0/K_0 , which is a measure of the adaptability of the signaling cascade, indicates that the main contribution of tunable thresholds is actually the appearance of a saddle point, which renders the extinction of the population stable, at intermediate P_0/K_0 values; as this control parameter increases the saddle point merges with the APC-limited state and both disappear. In this TAT model the only way the population of autoreactive T cells can persist is by competition for limited numbers of APCs; if TAT effects predominate the population will be extinct.

Discussion

The model offers a mechanism of self-tolerance by prevention of autoreactive T cell expansion, and their eventual deletion from the circulating pool. Therefore, the persistence of circulating anergic T cells requires their continuous influx from the thymus, as demonstrated by Sousa et al. (Sousa, 2003). This is not unreasonable because the thymus continuously produces T cells although the rate seems to decrease with age. However, when confronted with the Sakaguchi phenomenon the present model shows some important shortcomings.

Within the framework of our model, an adoptive transfer procedure in which lymphocytes are isolated *ex vivo* can be interpreted as an extra time-period during which the transferred lymphocytes remain free from the APCs. This manipulation would increase the responsiveness of lymphocytes as compared to the steady state *in vivo*, and this would be compatible with the experimental observations that CD4+CD25- lymphocytes from healthy subjects can induce autoimmunity in empty recipients. The fact that CD4+CD25+ T cells do not induce autoimmune disease in the recipient animals can also be interpreted by assuming that CD25+ cells have higher thresholds (higher values of phosphatase activity P) than CD25- cells. This is

not unreasonable because it is well documented that following anergy induction *in vitro* lymphocytes upregulate the CD25 molecule (Kuniyasu et al., 2000). However, the observation that cotransfers of the CD4+CD25⁻ and CD4+CD25⁺ result in tolerance cannot be easily interpreted under the light of this simplified model alone. As we have demonstrated "more cells should lead to more responsiveness or less anergy", therefore by adding CD25⁺ T cells to CD25⁻ cells one should be increasing the responsiveness of the CD25⁻ cells. The result of the co-transfers thus calls into action suppression mediated by regulatory T cells. Analysis of mathematical models of suppression by regulatory T cells suggested that the persistence and growth of the regulatory population is dependent on the autoreactive effector T cells they control, and that this dependency will increase the efficiency of suppressive function. This crosstalk between regulatory and effector cells can fully account for adoptive transfers of tolerance by CD4+CD25⁺ T cells, as well as for several other features of tolerance (Leon et al., 2003).

Tunable activation thresholds and suppression by regulatory T cells are not mutually exclusive mechanisms of self-tolerance. Already in their original proposal, Grossman & Paul (Grossman and Paul, 1992) suggested that anergic cells, with higher activation thresholds, could render naive cells anergic. Suppression by anergic cells has been shown *in vitro* (Lombardi et al., 1994; Taams et al., 1998), but the results are controversial (Kuniyasu et al., 2000), and the suppressive mechanism is unclear.

What dynamic properties are expected if T cell anergy is induced and maintained both by interactions with APCs and by interactions with other anergic T cells? We studied before (Leon et al., 2000) a model in which T_R cells convert T_E cells into the regulatory phenotype, showing that it has the same properties of a model in which T_R cells receive a growth factor from T_E cells. What would tunable activation threshold bring in addition to this? Consider the dependency of the fraction α of T_E and T_R cells activated upon conjugation on the frequency of conjugations and tuning. Consider also the frequency of T cell-APC interactions at the stable steady states of the (T, E) system with fixed α . Essentially the phase plane of the system will be maintained if the parameters are such that the frequency of conjugations with the APCs at the stable steady states are rare enough such that threshold tuning is not significant, i.e. α is practically constant; under these conditions, one would expect the extinction of both effector or regulatory T cells to be stable. However, if the parameters are such that the interactions with APCs are frequent enough to reduce the fraction of conjugated T_E and T_R that become activated then the steady states may disappear.

These additional complications of coupling suppression and anergy induction are non trivial and require proper modeling.

Methods

. Derivation of the quasi-steady state model of a single T-cell population:

The diagram in fig.1A can be translated into the following two differential equations and one conservation equation:

$$\frac{dT_F}{dt} = (1 + \alpha)dC - cT_F A_F - \delta \cdot T_F \quad (A1)$$

$$\frac{dC}{dt} = cT_F A_F - dC \quad (A2)$$

$$A = A_F + C \quad (A3)$$

where T_F is the density of free T cells, T_C is the density of APC-T-cell conjugates, A_F is the density of free APCs and A is the total density of APCs. The parameters are the rate constant of conjugate formation c , the rate constant of conjugate dissociation d , and the death rate constant δ . α is the probability that a T-cell is activated following activation. It depends on the internal state of the T-cell and it is defined according to eqn.10 in the main text, which uses the stationary probability density function of the phosphatase activity in the conjugated derived in Appendix B.

We are interested in following the total density of T cells in time denoted T :

$$T = T_F + C \quad (A4)$$

One practical reason to do this is that with the available experimental techniques it is very difficult to count free and conjugated T cells *in vivo*. Instead experimentalists isolated the mixture of T cells (conjugated or free) and counted them.

Taking the derivative of both sides of eqn. A4 we obtain:

$$\frac{dT}{dt} = \frac{dT_F}{dt} + \frac{dC}{dt} = d\alpha C - \delta T_F = d\alpha C - \delta(T - C) \quad (\text{A5})$$

which is to eqn.1 in the main text.

Assuming that the conjugates are in quasi-steady state we have:

$$\frac{dC}{dt} = cT_F A_F - dC = 0 \quad (\text{A6})$$

Substituting T_F and A_F by their expression in terms of A , T and T_C we obtain a second order equation:

$$c(T - C)(A - C) - dC = 0 \quad (\text{A6})$$

Solving it we obtain two solutions, a negative and a positive. Only the positive solution is physically meaningful and thus was considered and corresponds to eqn. 2 in the main text.

For phase-space and bifurcation analyses of this one-dimensional model we used the software Mathematica. The steady states were calculated numerically for each combination of parameters using the FindRoot routine of Mathematica, which implements the Newton method. The stability of each of these solutions was determined by linear stability analysis.

Appendix B. Derivation of the stationary distribution of phosphatase activity in a population of T cells:

The dynamics of probability density functions (PDFs) of the phosphatase activity in the subpopulations of conjugated and free T cells, respectively ρ_C and ρ_F , are described by the following set of first order partial differential equations:

$$\frac{\partial \rho_C}{\partial t} + \frac{\partial}{\partial P}(P_C \rho_C) = -d\rho_C + c_E \rho_F \quad (\text{B1})$$

$$\frac{\partial \rho_F}{\partial t} + \frac{\partial}{\partial P}(P_F \rho_F) = d\rho_C - c_E \rho_F \quad (\text{B2})$$

where P_C and P_F are the functions governing the dynamics of the phosphatase in the conjugated and free regimes (i.e. the right hand side of eqn. 4 with $\sigma>0$ and $\sigma=0$ respectively):

$$P_C = r_P(P_0(1 + \sigma) - P) \quad (B3)$$

$$P_F = r_P(P_0 - P) \quad (B4)$$

and $c_E=cA_F=c(A-C)$ and d are the per T cell transition rates from the free to the conjugated state and from the conjugated to the free state, respectively.

In search for the steady state solutions we make $\left(\frac{\partial P_C}{\partial t} = 0, \frac{\partial P_F}{\partial t} = 0\right)$ and

obtain the following set of ordinary differential equations:

$$\frac{\partial}{\partial P}(P_C \rho_C) = -d\rho_C + c_E \rho_F \quad (B5)$$

$$\frac{\partial}{\partial P}(P_F \rho_F) = d\rho_C - c_E \rho_F \quad (B6)$$

Noticing that the right hand sides of these two equations are symmetrical we can add them obtaining the following conservation:

$$\frac{\partial}{\partial P}(P_C \rho_C + P_F \rho_F) = 0 \quad (B7)$$

which upon integration leads to:

$$P_C \rho_C + P_F \rho_F = K \quad (B8)$$

Because ρ_C and ρ_F are PDFs, which cannot be negative, there must be at least one value P_1 such that $\rho_C(P_1) = \rho_F(P_1) = 0$. Therefore we have $K=0$ which leads to the following relation:

$$P_F \rho_F = -P_C \rho_C \quad (B9)$$

Solving this equation for ρ_F and substituting in eqn. B2 we obtain the following ordinary differential equation:

$$\frac{\partial \rho_C}{\partial P} = \left(-\frac{d}{P_C} + \frac{c_E}{P_F} - \frac{\partial P_C}{\partial P} \right) \rho_C \quad (B10)$$

The solution of this equation is:

$$\rho_C = N e^{\int \left(-\frac{d}{P_C} + \frac{c_E}{P_F} - \frac{\partial P_C}{\partial P} \right) \frac{1}{P_C} dP} \quad (B11)$$

Given the definitions of P_F and P_C according to eqns B3 and B4 the integrals are:

$$\int -\frac{d}{P_C} dP = \log |P_0(1+\sigma) - P|^{\frac{d}{r_P}} \quad (B12)$$

$$\int \frac{c_E}{P_F} dP = \log |P_0 - P|^{\frac{c_E}{r_P}} \quad (B13)$$

$$\int -\frac{\partial P_C}{\partial P} \frac{1}{P_C} dP = \log |P_0(1+\sigma) - P|^{-1} \quad (B14)$$

Replacing these integrals in eqn. B11 we get the following equation that corresponds to eqn. 9 in the main text.

$$\rho_C = \begin{cases} N|P_0(1+\sigma) - P|^{\frac{d}{r_p}-1}|P_0 - P|^{\frac{c_E}{r_p}}, & P_0 \leq P \leq P_0(1+\sigma) \\ 0, & \text{otherwise} \end{cases} \quad (\text{B15})$$

The solution is branched because at the steady state the values of P are always contained in the interval $[P_0, P_0(1+s)]$, whose extremes are the steady state values of P predicted according to eqn 4 in the main text if T cells would be either always free or always conjugated, respectively.

Acknowledgements

The authors are grateful to Jocelyne Demengeot, Zvi Grossman, António Coutinho, António Bandeira, Nuno Sepúlveda, and Íris Caramalho for many, many discussions on self-tolerance, and for their encouragement. This article is based on the Ph.D. thesis of Kalet Leon and João Sousa (available online at: <http://eao.igc.gulbenkian.pt/ti/index.html>). The work was financially supported by *Fundação para a Ciência e Tecnologia*: grants P/BIA/10094/1998, POCTI/36413/99, and POCTI/MGI/46477/2002; and fellowships to JF (Praxis/BCC/18972/98), JS (BD/13546/97), KL (SFRH/BPD/11575/2002), DM (SFRH/BD/2960/2000) and TP (SFRH/BD/10550/2002).

References

- Acha-Orbea, H. and H. R. MacDonald (1995). "Superantigens of mouse mammary tumor virus." *Annu Rev Immunol* **13**: 459-86.
- Barthlott, T., G. Kassiotis and B. Stockinger (2003). "T cell regulation as a side effect of homeostasis and competition." *J Exp Med* **197**(4): 451-60.
- Burnet, F. (1957). "A modification of Jerne's theory of antibody production using the concept of clonal selection." *Aust J Sci* **20**: 67-69.
- Clements, J. L., N. J. Boerth, J. R. Lee and G. A. Koretzky (1999). "Integration of T cell receptor-dependent signaling pathways by adapter proteins." *Annu Rev Immunol* **17**: 89-108.

- De Boer, R. J. and A. S. Perelson (1994). "T cell repertoires and competitive exclusion." J Theor Biol **169**(4): 375-90.
- Germain, R. N. and I. Stefanova (1999). "The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation." Annu Rev Immunol **17**: 467-522.
- Grossman, Z. and W. E. Paul (1992). "Adaptive cellular interactions in the immune system: the tunable activation threshold and the significance of subthreshold responses." Proc Natl Acad Sci U S A **89**(21): 10365-9.
- Grossman, Z. and W. E. Paul (2000). "Self-tolerance: context dependent tuning of T cell antigen recognition." Semin Immunol **12**(3): 197-203; discussion 257-344.
- Grossman, Z. and W. E. Paul (2001). "Autoreactivity, dynamic tuning and selectivity." Curr Opin Immunol **13**(6): 687-98.
- Kisielow, P., H. Bluthmann, et al. (1988). "Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4+8+ thymocytes." Nature **333**(6175): 742-6.
- Koshland, D. E., Jr., A. Goldbeter and J. B. Stock (1982). "Amplification and adaptation in regulatory and sensory systems." Science **217**(4556): 220-5.
- Kuniyasu, Y., T. Takahashi, et al. (2000). "Naturally anergic and suppressive CD25(+)CD4(+) T cells as a functionally and phenotypically distinct immunoregulatory T cell subpopulation." Int Immunol **12**(8): 1145-55.
- Langman, R. E. and M. Cohn (1987). "The E-T (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protection." Mol Immunol **24**(7): 675-97.
- Lederberg, J. (1959). "Genes and antibodies." Science **129**(3364): 1649-53.
- Leon, K., A. Lage and J. Carneiro (2003). "Tolerance and immunity in a mathematical model of T-cell mediated suppression." J Theor Biol **225**(1): 107-26.
- Leon, K., R. Perez, A. Lage and J. Carneiro (2000). "Modelling T-cell-mediated suppression dependent on interactions in multicellular conjugates." J Theor Biol **207**(2): 231-54.
- Lombardi, G., S. Sidhu, R. Batchelor and R. Lechler (1994). "Anergic T cells as suppressor cells in vitro." Science **264**(5165): 1587-9.
- Milutinovic, D., J. Carneiro, M. Athans and P. Lima (2004). "Modeling the hybrid dynamics of TCR signaling in immunological synapses." J Theor Biol **submitted**.
- Nicholson, L. B., A. C. Anderson and V. K. Kuchroo (2000). "Tuning T cell activation threshold and effector function with cross-reactive peptide ligands." Int Immunol **12**(2): 205-13.

- Pereira, P., L. Forni, et al. (1986). "Autonomous activation of B and T cells in antigen-free mice." Eur J Immunol **16**(6): 685-8.
- Polic, B., D. Kunkel, A. Scheffold and K. Rajewsky (2001). "How alpha beta T cells deal with induced TCR alpha ablation." Proc Natl Acad Sci U S A **98**(15): 8744-9.
- Sakaguchi, S., N. Sakaguchi, et al. (1995). "Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases." J Immunol **155**(3): 1151-64.
- Sakaguchi, S., N. Sakaguchi, et al. (2001). "Immunologic tolerance maintained by CD25+ CD4+ regulatory T cells: their common role in controlling autoimmunity, tumor immunity, and transplantation tolerance." Immunol Rev **182**: 18-32.
- Singh, N. J. and R. H. Schwartz (2003). "The strength of persistent antigenic stimulation modulates adaptive tolerance in peripheral CD4+ T cells." J Exp Med **198**(7): 1107-17.
- Smith, K., B. Seddon, et al. (2001). "Sensory adaptation in naive peripheral CD4 T cells." J Exp Med **194**(9): 1253-61.
- Sousa, J. (2003). Modeling the antigen and cytokine receptors signalling processes and their propagation to lymphocyte population dynamics. Lisbon, University of Lisbon: 161.
- Stefanova, I., B. Hemmer, et al. (2003). "TCR ligand discrimination is enforced by competing ERK positive and SHP-1 negative feedback pathways." Nat Immunol **4**(3): 248-54.
- Taams, L. S., A. J. van Rensen, et al. (1998). "Anergic T cells actively suppress T cell responses via the antigen-presenting cell." Eur J Immunol **28**(9): 2902-12.
- Tanchot, C., D. L. Barber, L. Chiodetti and R. H. Schwartz (2001). "Adaptive tolerance of CD4+ T cells in vivo: multiple thresholds in response to a constant level of antigen presentation." J Immunol **167**(4): 2030-9.
- Witherden, D., N. van Oers, et al. (2000). "Tetracycline-controllable selection of CD4(+) T cells: half-life and survival signals in the absence of major histocompatibility complex class II molecules." J Exp Med **191**(2): 355-64.
- Wong, P., G. M. Barton, K. A. Forbush and A. Y. Rudensky (2001). "Dynamic tuning of T cell reactivity by self-peptide-major histocompatibility complex ligands." J Exp Med **193**(10): 1179-87.

VII

Discussion

This thesis addressed the origin and implications of non-genetic individuality brought upon by stochastic molecular events in single cells. We critically analyzed existent models for several distinct types of population heterogeneity generated by stochastic events and proposed different models that addressed the outlined difficulties. Furthermore, we explored some of the consequences of this heterogeneity in two different kinetic regimes (static and dynamic variation) and showed how this heterogeneity can be relevant for the response of a population as well as for the reinterpretation of results obtained using traditional techniques. Specifically, in chapter II we identified difficulties that previous models faced when trying to explain the specific lognormal shape that is found to be ubiquitous in protein copy number profiles. We proposed that multiplicative noise within several cell processes, which can be viewed as multistep extensions to the central dogma, underlies the nature of this distribution. In chapter III, we elucidate the mechanism by which monoallelic expression takes place by assessing the plausibility of several popular mechanisms proposed in the literature. By mathematically formalizing such mechanisms we show that several of these are not plausible and propose a general model of transcription regulation that has is compatible with the observed data. These two chapters conclude our investigations on the origin of these two types of heterogeneity, the fluctuations in protein copy numbers and in locus competence. Both explanations make use of extensions to the central dogma and intra-cellular stochastic events in order to be able to explain the observed non-genetic individuality. These previous works have established a notion of stochastic individuality, which in turn evokes a shift in point of view from the single cell to a population whose constituents have distinct characteristics. This prompted us to investigate the implications of the somatic variation induced by stochastic events for the response of a population, both in terms of how it translates to heterogeneity of the population and as a hindrance for some modeling techniques. We addressed these matters from two dynamic limits: when the fluctuations of relevant proteins are slow in relation to the process being studied and when these fluctuations are in the same timescale as the cell division process. As such, we next turned our attention to the propagation of this heterogeneity to the population level. In chapter IV, we assess the impact of distributed protein copy numbers on several networks motifs within the cell. Here we make the explicit assumption that the dynamics of such networks are faster than the

dynamics of the processes that give rise to this heterogeneity, hence this investigation pertains what we call “static heterogeneity”. We show how different modules transform their inputs due to their intrinsic heterogeneity and how such heterogeneity propagates in these networks, paving the way for an understanding of networks as systems whose components have different levels across the population. Moreover, we conclude that observing just the average value of such populations can mask the dynamics of the single cell due to the distributed nature of the population. We then, in chapter V, address the question of a possible adaptive role for somatic variation by copy number fluctuations, concluding that the structure of these fluctuations is relevant for the adaptation dynamics of the population.

In chapter VI, we couple one of the circuits of chapter IV to extracellular signals derived from interactions with another population. The stochastic nature of these extracellular interactions creates an heterogeneous population in their ability to proliferate. This hints at how a multicellular organism can make use of this source of non-genetic variation to regulate its constituent populations.

In this thesis we argued that stochastic effects inherent to the cells metabolism generates non-genetic individuality in an isogenic population. Moreover, we investigated the implications of such heterogeneity at the level of signal transduction, gene expression and population dynamics. We tried to convey a new way to look at cell populations and their relation to their environment based on the stochastic events at the single cell level. This view represents an emerging perspective in which determinism is based not on the dynamics of the single cell but on the evolution of a distribution of probabilities, in which each cell is a realization of a stochastic process. Within this perspective, programs of differentiation, cellular identity and individuality acquire new meanings.

As the cell is a dynamical unit, constantly reacting to the environment and whose parts are constantly being renewed, so a population of cells can be seen as an entity whose parts are constantly changing and, through that process, constantly readjusting to the environment. In fact, as we have seen on chapter V and VI, the environment can be a reorganizing force in the population, shaping and maintaining the heterogeneity within the population. Moreover, this can be extended to higher levels of biological organization, from multicellular organisms to ecosystems, as discussed in the rest of this discussion.

Although each chapter had its own specific discussion, several broader questions relating to those works have not yet been addressed. In the next sections we will discuss these broader questions, relating them to our work and to related approaches in the literature.

Timescales and the intrinsic-extrinsic noise question

Variability of protein copy numbers within cell populations, usually termed “noise” in the literature, was proposed to be further decomposed in two components: intrinsic and extrinsic noise (Swain et al., 2002). This classification has been motivated by a strict adherence to the simplest extension of the central dogma: that stochastic effects at the level of transcription and translation, brought upon by the finite number of molecules involved, were the major cause of variability. In this sense, “intrinsic” variability was defined as the variability attributable to these processes and “extrinsic” variability would mean variability introduced by all other processes that might influence the expression of the gene, such as stochasticity in signal transduction or variability in molecules not directly linked to the gene. These two sources of heterogeneity were proposed to be distinguishable by measuring protein expression on two different alleles controlled by the same promoter (Raser and O’Shea, 2004). Specifically, the authors constructed a yeast strain in which the transcription of two different fluorescent proteins (CFP and YFP) is controlled by identical promoters. Given the assumption that extrinsic noise is unrelated to the gene in question, and hence constant in one cell, the difference between the expression level of the two alleles would cancel the extrinsic component of the variability (Swain et al., 2002), leaving only the intrinsic component. As such, and since the data thus obtained was highly correlated, manifesting as a ellipsoid whose major axis was along the diagonal, the authors then propose that extrinsic noise can be quantified by the variance along such line and intrinsic noise as the variance along the orthogonal direction. Formally, these two variances are related to the eigenvalues of the covariance matrix for the data, although the authors present different formulae for these variances. Thus, the authors imply that deviations from correlated expression from the two alleles is attributable to “intrinsic” causes while heterogeneity in level of expression in correlated expression is attributable to “extrinsic” causes. This analysis found that “extrinsic” variability is orders of magnitude higher than “intrinsic” noise.

This experimental system is equivalent to the one we used when assessing the nature of the mechanism that gave rise to monoallelic expression in T cells in chapter III: two distinguishable alleles controlled by identical regulatory regions. However, these data present no correlations, and they are not shaped as an ellipsoid. How could the previous analysis be carried out on that data? Surely that, according to the previous definition, the IL-10 allelic expression data shows significant amount of

“extrinsic” noise and none of that is apparent if we apply the above methodology to it. In fact, by calculating the eigenvectors on these data, the major eigenvalue is the one associated with the anti diagonal eigenvector, which, under the previous framework, is interpreted as meaning that the major source of variability is intrinsic in nature.

The notion of intrinsic and extrinsic noise was widely widespread in the literature (Swain et al., 2002; Raser and O'Shea, 2004; Newman et al., 2006; Scott et al., 2006; Sigal et al., 2006; Volfson et al., 2006). However, since within our framework all heterogeneity is explained as the result of the same process, we disagree with these definitions for intrinsic and extrinsic noise and even with the distinction itself. We find that there is really no distinction between the two and that, at most, it is operationally more relevant to distinguish between timescales of processes. As such, a useful distinction between processes would be one based on their relative timescales. We would call “noise” to processes whose stochastic dynamics are fast compared to an observation process, meaning that each time a specific observation took place, observations of that variable are uncorrelated. We would reserve the term “heterogeneity” where significant correlation persists between measurements. It should be noted that, according to our models, there is really no distinction between the two, since the former is the basis for the latter. It is essentially a matter of complexity of the generative process. Consistently, we would restrict the term “extrinsic” to exogenous processes of generation of diversity, such as interactions with other populations, as we show in chapter VI. In this context, within our work, we have been mainly concerned with extrinsic heterogeneity.

In fact, within the framework of the work developed in chapters II and III, the data by Raser *et al* could easily be explained by what they call extrinsic sources of heterogeneity. In chapter III, we proposed a general model for the regulation of transcription by epigenetic modification. In this model, this regulation was achieved by having multiple intermediate states of locus competence mediated by the presence of modifications to the locus. The high correlation of expression from both alleles would mean that the distribution of these modifications would be highly skewed for the “open” state, while the heterogeneity in level of expression would be attributable to posttranslational events or heterogeneity in signal transduction due to variability in intervenient molecules (according to chapter II). As such, we would agree with previous authors that essentially all variability is due to what was called “extrinsic” noise, and we would extend that in saying that not only the cells are heterogeneous but also that such heterogeneity is highly heritable.

In fact, many works have reiterated that extrinsic variability was responsible for the heterogeneity in cell populations. In this context, it is assumed that the heterogeneity observed in a given process is essentially influenced by factors that are themselves distributed across the population but that are not part of the process under observation. This seems to be an infinite regression as those “extrinsic” factors, if under scrutiny, would themselves be influenced by other “extrinsic” factors. As it is apparent from our work in chapter II we closed this infinite regression by noting that such heterogeneity in these “extrinsic” factors (in this case, our catalysts) are themselves created by a similar process. In fact, it seems that the structure of the cellular metabolism creates an environment that sustains the noise and leads to individuality, with its dense network of interactions making every cellular component dependent on many others.

Inheritance and Memory

Individuality is intimately linked to the question of heritability. In the work developed in the first two chapters, we identify two modes of inheritance for the two sources of individuality analyzed. Inheritance of locus modifications depends upon the correct copying of these during cell division. The specific mechanism of epigenetic inheritance is not yet fully understood. The two epigenetic locus modifications known are DNA methylation and histone post-translational modifications. DNA methylation is associated with repressible gene transcription and is believed to be irreversible (Martin and Zhang, 2007). As such, only cell division can generate hemimethylated DNA which is the preferential substrate for methyltransferases which copy methylation states from one strand to the other, perpetuating the epigenetic information. As has been discussed elsewhere (Lim and van Oudenaarden, 2007), the stable inheritance of this depends on the high efficiency of these factors, leading to errors when this information transfer is not perfect. On the other hand, the mode of inheritance of histone modifications is still elusive (Martin and Zhang, 2007). However, due to the nature of these modifications, it is expected that cell cycle to be important for this inheritance. In fact, it has been reported that cell division is an important event in differentiation events that involve epigenetic modifications, such as in Th1/Th2 differentiation in T cells (Ansel et al., 2003).

Inheritance of somatic variation in protein copy number is possible through equipartitioning of protein copies at mitosis, leading to the notion of cytoplasmatic inheritance. The stable inheritance of this phenotype depends on the high number of

molecules present as partitioning errors follow a poisson distribution and hence its variance acquires less importance as the average value is increased. In Sigal et al (Sigal et al., 2006), where fluctuations of specific proteins are analyzed at the single cell level, it is shown that, although protein levels are constantly fluctuating, cell division does not have an impact on their autocorrelation, indicating that these is in fact a stable inheritance of this phenotype.

Control of stochastic effects

One question that seems to pervade the literature on the subject of stochastic gene expression regards the control of stochasticity (Ozbudak et al., 2002; Raser and O'Shea, 2004; Suel et al., 2007). This concern stems directly from the notion that stochastic effects are intrinsically deleterious for the proper functioning of the cell. This stems from the informational view of the cell that stemmed from the enunciation of the central dogma: "Information only flows from DNA to RNA and then to protein" (although the meaning referred to sequence information). Faithful decoding of this information would require any "noise" to be minimized. Indeed, it is true that some cellular events require precision in their regulation: cell cycle phases and apoptosis come to mind. It is clear that certain phases of the cell cycle, especially within mitosis, cannot be made until the previous is completed. Cell cycle checkpoints, however, depend on protein concentrations in order to correctly operate. It seems puzzling, then that a cell can cope with heterogeneity in these situations. However, as we've seen on chapter IV, switching mechanisms can be resilient to this variability, its heterogeneity manifesting itself essentially on the time it takes for the switch to take place, as long as the signal is properly chosen. In fact, recent works (Doncic et al., 2006) analyze this question in relation to the spindle checkpoint and conclude that indeed it is resilient to dynamical perturbations.

In fact, it was observed that genes related to the same module or function presented significant correlation on their expression levels (Sigal et al., 2006), hence diminishing the effects of heterogeneity by preserving magnitude relations between related molecules.

This seems to be the most parsimonious way to control the effects of heterogeneity since, according to our work on chapter II, a mechanism to control such heterogeneity would be rather expensive to implement. For example, in our proposed mechanism for the functional expression of proteins, a mechanism of end-point inhibition, which has been argued to confer some advantages in unbranched

pathways (Alves and Savageau, 2000), seems implausible as an effective way to control the variance of these proteins. This because decorrelation between the elements of the chain happens quickly and, for any reasonable sized chain, the last element is essentially decorrelated from the first, rendering a long-range feedback mechanism useless. In fact, under certain conditions that this mechanism would just increase the overall variance (unpublished results), as corrections to the first element by the mechanism would be akin to an extra stochastic term feeding in that element.

A distinction should be made between variance control and noise resistance; while variance control pertains to the control the variance of expression of one protein, noise resistance is the property of systems to robustly endure rapid variations on a signal input or in one of its components. Most works on control of stochastic effects have focused on the latter (Ozbudak et al., 2002; Eissing et al., 2005; Pedraza and van Oudenaarden, 2005; Chen and Wang, 2006; Doncic et al., 2006; El-Samad and Khammash, 2006). However, constitutive molecules with average copy numbers of the order of 10^4 , such as the TCR and many other proteins, are unlikely to be prone to significant fast stochastic variation. As such, and since we were more concerned with the role of the individual somatic variation, we have focused on the former in chapter IV, where we analyzed the sensitivity to parameter variation (what we called static heterogeneity) in both steady states and dynamics of several common network motifs. We have shown how the population's heterogeneity is mapped to a distributed output, realizing that the structure of interactions determines the outputs variance. For example, in the last model analyzed in that chapter, although the critical signal strength is distributed in the population and is fairly sensitive to the heterogeneity in total protein amount of R2, the higher stable steady state is nearly insensitive to this heterogeneity. As such, mechanisms such as this could provide a way to control the variance, in situations where it is required, as slow variations of these molecules can be faithfully described by the bifurcation diagrams presented at that chapter.

The lognormal distribution

We made a thorough investigation on the origin and implications of the lognormal distribution of protein copy numbers. Although other works ascribe different origins for this distribution other than the one we described in this paper (Furusawa et al., 2005; Krishna et al., 2005), several facts seem to favour our own, as discussed in chapter II. In fact, we've shown that many ubiquitous processes within the cell lead to

a lognormal distribution of their products. The ubiquity of such distributions raises the possibility that this shape might be advantageous for a cell population. In fact, we have shown in chapter V that, when compared to other distributions, the lognormal distribution entails very different characteristics for the somatic adaptation of a population to a new environment than other distributions. This wide flexibility of adaptation entailed by the lognormal distribution can be argued to allow increased control over the dynamics of adaptation. Moreover, in chapter II we've seen the flexibility of a switching mechanism (akin to one of the examples described on chapter IV) whose thresholds are lognormal distributed in regulating the percentage of cells that respond to a stimulus. In fact, the versatility of this distribution allows very specific control over the percentage of responding cells.

One of the ways to further understand the putative adaptive value of a certain trait and make a case for selective pressure towards that trait is to understand the mechanism that gives rise to it. Our generative mechanism for this particular shape involves multiplicative noise brought upon by the cell's organization. Two alternative points of view on this are possible: multiplicative processes might have been selected for, under a selection pressure to have multiplicative noise (with its long tails and dynamics) or other pressures might have given rise to multiplicative processes and the lognormal distribution is just an inevitable byproduct of this. However, this latter view doesn't mean that the observed lognormal distribution in protein copy numbers is without function, as the cells might still have made use of it as it was a trait that was present. It is not hard to argue that populations within a multicellular organism make use of such distribution in their interactions with other populations. In fact, when two distributed populations interact (in direct cell-cell interactions), the stimulus of one population can be seen as a distributed input. And as can be seen on the first example of chapter IV – the simplest signaling network –, a distributed input would entail a lognormal distributed output with a shape parameter equal to the sum of both receptor and stimulus shape parameters (as the multiplication of two lognormal distributed variables is a lognormal with these parameters summed), hence displaying even heavier tails. This means that the range of representation of inputs is increased in the responder population, which raises the possibility of further refinement for alternate fates by additional circuitry. Additionally, many signals are sensed through multimeric receptors that assemble only in the presence of its ligand (e.g. the cytokine receptors) and, moreover, some chains are shared between several ligands. Since each chain is expressed independently, the resulting functional receptor will have a highly increased skewness across the population.

As such, the fact that many proteins are distributed as a lognormal in a cell population can have a profound impact in population responses by entailing long tails in the sensitivity to a stimulus.

The adaptive role of heterogeneity generated by somatic variation

Ever since the recognition of non-genetic individuality in cell populations that the possibility of a putative adaptive role of such phenomena has been discussed (Spudich and Koshland, 1976). Such questions are obviously linked to the question of how such heterogeneity is brought upon. Since it is now apparent that much of this non-genetic individuality is attributable to stochastic effects inherent to the cell's organization, an evolvability scenario for this phenomenon would, from a naïve point of view, seem unlikely. In fact, since stochastic effects are by definition unpredictable, they would interfere with the proper functioning of the cell, its sensing of the environment and response to particular stimuli. However, the discovery of phenomena with obvious adaptive value, such as bacterial persistence upon antibiotic exposure (Bigger, 1944), which has been shown to be partly induced by a noise-induced phenotypic switch (Balaban et al., 2004), cast some doubts over this point of view. In fact, it has been known for a long time that many prokaryotic species exhibit phase variation (Lederberg and lino, 1956), although only recently the connection to stochastic gene expression was made, at least in some cases (Arkin et al., 1998). It is clear now that bacterial populations maintain a standing non-genetic variation of different “differentiation” states, suggesting that these populations would rely on stochastic effects to generate such diversity. This, in turn, raises the possibility that bacterial populations are making use of this spontaneous diversity to better cope with a fluctuating environment, as an alternative to a purely sensorial regulated strategy. In fact, the putative adaptive role of such strategy was investigated and confirmed by several works (Balaban et al., 2004; Thattai and van Oudenaarden, 2004; Kussell and Leibler, 2005; Wolf et al., 2005; Wolf et al., 2005). All these works focused essentially on the adaptive role of having stochastic switching rates between discrete states of a population, in the context of a fluctuating environment, and addressed questions such as optimal switching rates and diversification strategies. However, the majority of these works focused in these switching rates disregarding the details of the underlying mechanism that gave rise to such switching. As such, it was generally assumed that all transitions were equally probable or that a strict hierarchy between these states existed and only nearest

neighbor transitions occurred. This is consistent with an extremely noisy view of the cell, in which switching mechanisms operate at extremely low levels of intervening molecules or at the brink of multiple bifurcations.

An obvious question regarding the evolvability by natural selection of this particular type of somatic variation, fluctuation structure of protein copy numbers, is that each individual cell is a realization of a stochastic process and could, at most, be selected by the particular amount of protein it exhibits at any given time. As such, it would be hard to select for a steady state distribution or fluctuation rate based on the individual cell. However, as we have seen on chapter II, the population reflects the whole breadth of the distribution and its response can be dependent on the population's particular fluctuation structure, as seen on chapters II, IV and V. This means that the population is a snapshot of all the possibilities and trajectories of the single cell, i.e. a phenotype induced by a particular genome. As such, selection can act at the level of the response of the population and the mechanisms of generation of somatic variation become selectable traits.

In chapter V we addressed the question of an adaptive role for dynamic heterogeneity within the context of fluctuations of one molecule, assuming a specific, albeit simple, mechanism and a deterministic environment. We assumed that protein levels had a continuous effect on the probability of division of individual cells and have shown that such population attains a new equilibrium with the selective pressure of the imposed environment. We have shown also that different fluctuation characteristics entail different kinetics of adaptation with some of these displaying a runaway adaptation. Again, it seems unlikely that a population that exhibits such traits would not evolve some mechanisms to take advantage of it.

For example, it is indeed entirely possible that some observations of upregulation of genes under a given stimulus do not involve specific regulation but indeed just relies on the natural heterogeneity and dynamical fluctuations of those gene products. This is indeed the old “nature or nurture” question under a new guise, as it is not just that the observed new population could be a an expansion of some preexistent minority population but that it could be induced by the selective environment and the fluctuation characteristics of the selected protein. Reciprocally, it is common to observe that populations isolated from animal models loose expression of specific proteins when cultured. This is often attributed to the absence of some essential factor in the culture medium. Our work suggests an alternative, in which what is lacking is the selective proliferation of individuals that express higher levels of a certain receptor – a subtle but important distinction.

Although for unicellular organisms it is hard to imagine other advantages for this diversification other than coping with a fluctuating environment (including immune evasion, which is related), since they are typically regarded as asocial organisms, for a multicellular organism many more advantages in distributed populations can be argued. The diversification strategies employed by prokaryotic populations in order to respond to stimulus could very well be put to use in the context of a multicellular organism. A multicellular organism establishes an environment for its constituent lineages, which is altered as the organism is challenged by environmental changes. These changes, as the induction of a factor for proliferation on a specific population (our case study), could act on a non-differentiated population that, through leaky transcription, some of its elements would express residual quantities of the factor of interest. This would lead to its selection and subsequent upregulation. This would also entail that this population would return to its undifferentiated state as these conditions disappear, unless additional regulation is present to fixate these changes. In contrast with a population making use of its intrinsic heterogeneity to better cope with a stochastic environment, in chapter VI we have shown how a stochastic environment can itself lead to an heterogeneity of protein levels in a population. This heterogeneity then leads to a non-trivial homeostatic regulation mechanism.

Perspectives and final remarks

This thesis tried to convey a new perspective on how a cell population or lineage makes use of the historical contingencies that grant individuality to its constituent single cells. This perspective might prove useful to interpret many different phenomena, especially within the complex organization of multicellular organisms. Indeed, many different fields are touched by this new view of the cell and the nature of cellular individuality. We have argued that the reality of non-genetic individuality calls for a shift in point of view from the single cell to the isogenic cell population as an environment assessing unit. In fact, as we have shown in chapters V and VI, the environment has a considerable impact on the maintenance and shaping of the population heterogeneity. As such, the internal structure of the population reflects both the internal dynamics of the cell but also the impact of the environment on it. Consider, for example the following mechanism: we have shown protein copy number heterogeneity allows that each cell has a distinct threshold of response to a given stimulus. This fact allows that, given a uniformly valued stimulus only a part of the population responds. If the action of response decreases the value of the

stimulus this allows for a simple rheostat model of response. This population rheostat could be in use in many disparate organizational levels, from immunology, where the action of lymphocytes indeed eliminates the signal, to division of labor in social insects, as it was previously proposed (Fewell, 2003). Many other regulation mechanisms might prove to be dependent on the somatic variation of individual cells.

In immunology, many different aspects are touched by this new perspective. Selection at the thymus, for example, has been known to be sensitive to the amount of signaling each thymocyte senses. Moreover, different fates have been hypothesized to be determined by this “signal strength” (the differential avidity model), extremely low or high signals leading to death, while several degrees of intermediate strength would lead to conventional T cells, CD4+ and CD8+, or regulatory T cells. Considering that avidity is a function of both affinity and amount of receptor, how could that work when cell does not have a definite TCR copy number? This suggests that the interpretation of the signal is robust enough so that differences in TCR concentration do not matter (meaning that the thresholds for the decisions should be separated by more than the typical variance of the whole signaling cascade). Maybe it is possible for the whole process to take long enough such that each thymocyte has enough time to fluctuate over the whole range of its associated distribution, allowing the selection process to sample the whole breadth of these fluctuations.

A mechanism such as the stochastic population rheostat described above makes sense to be operating in the immune system, as this is the prototypical complex system that responds and adapts to its environment. In fact, as “naïve” CD4+ T cells become activated by exposure to antigen loaded APCs, they can differentiate into two alternative fates, Th1 and Th2, distinguishable by the cytokine profile they secrete (their hallmarks being Interferon gamma and interleukin-4, respectively). However, it has been shown that, when populations of naïve cells are first stimulated, they express broad spectrums of these cytokines, with no definite Th identity (Grogan et al., 2001). This identity seems to be established by further cytokine signaling, driving the expression of a pair of mutual repressing transcription factors, GATA-3 and T-bet. It seems that this process requires cell division and leads to epigenetic changes at the relevant cytokine locus (Ansel et al., 2003; Chang and Aune, 2007). The mutual repression of GATA-3 and T-bet most likely creates a bistable system (Hofer et al., 2002), which should lead to the establishment of the Th identity of the cell. In light of the work developed in this thesis, the proneness of some cells to be directed to one of these fates could be linked to different copy

numbers of these transcription factors or of the signal transduction machinery. Moreover, the requirement of epigenetic modification for this differentiation to be permanent could be linked to the fluctuations of these transcription factors, making this decision unreliable. As such, we would expect that indeed, in a general stimulation environment, there would be coexistence of both phenotypes, and that selection of these through differential proliferation induced by other feedback mechanisms would polarize the population.

Previous work has also considered the sharing of the gamma chain on the IL-4 and IL-2 receptors (Sousa, 2003). In this work, it is argued that the amount of this gamma chain can modulate the levels of Th2 commitment and the maintenance of precursors of a given clone. As can be seen on figure 1 of chapter II, a polyclonal staining for this receptor chain shows an extremely broad distribution. This suggests, in light of our work, that this broad distribution is used to achieve fine control over the percentage of cells that are committed to the Th2 fate.

Regulatory T cells have been under immense scrutiny ever since their discovery. Although initially thought to be one thymus-derived population, enriched in the CD4⁺CD25⁺CD45Rb^{low} subpopulation, the so called natural regulatory T cells, it was soon discovered that several other populations could exhibit the same regulatory properties. The field was soon swamped with reports of regulatory cells that did not express the classical markers and did not follow the standard developmental pathway (i.e. thymus derived). The major distinction from these classical (naturally occurring) Tregs and the alternative variety seems to be the constitutive expression of FOXP-3 by the former, a transcription factor. It seems that some of these cells are induced in the periphery, under certain conditions. Considerable efforts have been made to identify the specific factors that induce the regulatory phenotype in normal T cells. However, our work suggests that perhaps no such factor exists. Indeed, is it possible that these induced regulatory cells are induced from the normal T cell pool and are just the fraction of these that express a higher (or lower) amount of certain proteins at the moment of induction. Since these cells seem to be responsible for the control of ongoing immune reactions, this seems like a straightforward mechanism to regulate the number of cells that are recruited to this fate, as this number would be indexed to the number of cells recruited to the immune reaction, due to their standing variation.

Cell lineages are the hallmark of multicellular organisms. These are established during development, through intercellular signaling processes, and perpetuated throughout the organism's life span. Although these decisions seem to be extremely

regulated, even the simpler organisms, such as the nematodes, in which the fate of every cell is mapped and shows amazing reliability, show some degree of stochasticity in some of these decisions. In particular, the Ventral Cell/Anchor Uterus decision relies on stochasticity to be made (Sternberg and Felix, 1997). In higher organisms, in which the complexity is higher, it is plausible that the frequency of usage of this type of mechanism to take developmental decisions is much higher. Moreover, the parallels of cell lineage/differentiation with the observed prokaryotic heterogeneity of “differentiation” states suggests a common origin. It is possible that lineage evolution might have primarily relied on the individual somatic variation and developed mechanism to freeze that fate, later on. In fact, this non-genetic individuality in general seems to be a way for genomes to explore on a finer scale the fitness landscape, enabling them to easily traverse fitness valleys, as they reduce the force of selection on the genome itself.

The mechanism of action of morphogen gradients in developmental fate decision has been elusive for a long time. One of the major mysteries is how two to three fold differences in morphogen concentration are able to induce different fate decisions, when the signaling pathways seem to be linear (as changes in morphogen concentration induce similar changes in its transcriptional targets) and only one morphogen is present. This is even more so when several of this differentiation decisions must happen for the same morphogen gradient (Ashe and Briscoe, 2006). It is possible that the answer to such questions is that each cell senses the morphogen gradient differently, due to their inherent variability, and then a migration of the population takes place, making use of the gradient for the correct placing of the cell. This entails that there will be correlations between the distance to the source of the gradient and the fate of the cell (as observed) but dispenses with complex mechanisms of differentiation and extremely accurate gradient sensing. Another caveat in morphogen gradient models has been that these gradients decay exponentially, although cells can sense them through many length scales. Again, the long tails of lognormal distributions might prove useful in making sure that some cells are able to sense the low concentrations (very far from the maximum of the gradient).

Other examples can be found in neuron networks have been described as Bayesian inference modules, calculating probabilities of events. It has been argued that neurons are actually calculating posterior distribution of events given a (perceptive) stimulus (Pouget et al., 2003). The major problem with this approach has been to understand how neurons encode and decode distributions from a mechanistic point of view. The framework described in this thesis allows an obvious possibility. In fact,

the fluctuation structure of relevant proteins could be somatically tuned in order to have an ensemble of neurons encoding distributions. In fact, neuron sensitivities have been described to be distributed in a population of neurons and to be transient, i.e. the sensitivity of a neuron changes in time and if you isolate a subpopulation it reconstitutes the full distribution of sensitivities (Park et al., 1990). The distribution of these sensitivities is wide and seems to be shaped as a lognormal distribution raising the hypothesis that these sensitivities are controlled by the amount of one or more lognormal distributed proteins present in these neurons.

Moreover, behavioral variability in organisms with a central nervous system has been attributed to residual noise, a by-product of a complex brain (even giving rise to the semiserious Harvard Law of Animal Behavior (Grobstein, 1994): "Under carefully controlled experimental conditions, an animal will behave as it damned well pleases"). However, recent works in flies (Maye et al., 2007) have provided evidence that erratic behavior (e.g. during mating chases) is indeed spontaneous and not due to environmental fluctuations. Moreover, they've shown that it is at least partly due to stochastic effects, although with long range correlations. Consistently with this view of the brain, the long-range correlations being consistent with slow fluctuations of proteins in networks of neurons and spontaneous behavior being compatible with a probabilistic encoding of information on neurons. It is indeed possible that "voluntariness" ("free will") might be a consequence of stochastic effects on one's brain.

All these speculative examples illustrate an novel way of looking at the organization and regulation of the multicellular organism or system, from the perspective of somatic variation of individual cells and the heterogeneity that that leads to, at the population level. Much work, both theoretical and experimental, needs to be done in order to further investigate the implications of heterogeneity in social interactions. We hope to have given a stimulating argument and to have provided a food for thought on a reappraisal on the nature of individuality, populations and individual based selection.

References

Alves, R. and M. A. Savageau (2000). "Effect of overall feedback inhibition in unbranched biosynthetic pathways." Biophys J **79**(5): 2290-304.

- Ansel, K. M., D. U. Lee and A. Rao (2003). "An epigenetic view of helper T cell differentiation." Nat Immunol **4**(7): 616-23.
- Arkin, A., J. Ross and H. H. McAdams (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." Genetics **149**(4): 1633-48.
- Ashe, H. L. and J. Briscoe (2006). "The interpretation of morphogen gradients." Development **133**(3): 385-94.
- Balaban, N. Q., J. Merrin, et al. (2004). "Bacterial persistence as a phenotypic switch." Science **305**(5690): 1622-5.
- Bigger, W. (1944). "Treatment Of Staphylococcal Infections With Penicillin by Intermittent Sterilisation." The Lancet **244**(6320): 497-500.
- Chang, S. and T. M. Aune (2007). "Dynamic changes in histone-methylation 'marks' across the locus encoding interferon-gamma during the differentiation of T helper type 2 cells." Nat Immunol.
- Chen, B. S. and Y. C. Wang (2006). "On the attenuation and amplification of molecular noise in genetic regulatory networks." BMC Bioinformatics **7**: 52.
- Doncic, A., E. Ben-Jacob and N. Barkai (2006). "Noise resistance in the spindle assembly checkpoint." Mol Syst Biol **2**: 2006 0027.
- Eissing, T., F. Allgower and E. Bullinger (2005). "Robustness properties of apoptosis models with respect to parameter variations and intrinsic noise." Syst Biol (Stevenage) **152**(4): 221-8.
- El-Samad, H. and M. Khammash (2006). "Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks." Biophys J **90**(10): 3749-61.
- Fewell, J. H. (2003). "Social insect networks." Science **301**(5641): 1867-70.
- Furusawa, C., T. Susuki, et al. (2005). "Ubiquity of Log-normal Distributions in Intracellular Reaction Dynamics." Biophysics **1**: 25.
- Grobstein, P. (1994). Variability in Brain Function and Behavior. The Encyclopedia of Human Behavior. V. S. Ramachandran, Academic Press. **4**: 447-458.
- Grogan, J. L., M. Mohrs, et al. (2001). "Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets." Immunity **14**(3): 205-15.
- Hofer, T., H. Nathansen, et al. (2002). "GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model." Proc Natl Acad Sci U S A **99**(14): 9364-8.
- Krishna, S., B. Banerjee, T. V. Ramakrishnan and G. V. Shivashankar (2005). "Stochastic simulations of the origins and implications of long-tailed distributions in gene expression." Proc Natl Acad Sci U S A **102**(13): 4771-6.

- Kussell, E. and S. Leibler (2005). "Phenotypic diversity, population growth, and information in fluctuating environments." Science **309**(5743): 2075-8.
- Lederberg, J. and T. Iino (1956). "Phase Variation in Salmonella." Genetics **41**(5): 743-57.
- Lim, H. N. and A. van Oudenaarden (2007). "A multistep epigenetic switch enables the stable inheritance of DNA methylation states." Nat Genet **39**(2): 269-75.
- Martin, C. and Y. Zhang (2007). "Mechanisms of epigenetic inheritance." Curr Opin Cell Biol.
- Maye, A., C. H. Hsieh, G. Sugihara and B. Brembs (2007). "Order in spontaneous behavior." PLoS ONE **2**: e443.
- Newman, J. R., S. Ghaemmaghami, et al. (2006). "Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise." Nature **441**(7095): 840-6.
- Ozbudak, E. M., M. Thattai, et al. (2002). "Regulation of noise in the expression of a single gene." Nat Genet **31**(1): 69-73.
- Park, S., J. Principe and J. Smith (1990). "Tdat - time domain analysis tool for eeg analysis." IEEE Trans. Biomedical Engineering **37**(8): 803-811.
- Pedraza, J. M. and A. van Oudenaarden (2005). "Noise propagation in gene networks." Science **307**(5717): 1965-9.
- Pouget, A., P. Dayan and R. S. Zemel (2003). "Inference and computation with population codes." Annu Rev Neurosci **26**: 381-410.
- Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-4.
- Scott, M., B. Ingalls and M. Kaern (2006). "Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks." Chaos **16**(2): 026107.
- Sigal, A., R. Milo, et al. (2006). "Variability and memory of protein levels in human cells." Nature **444**(7119): 643-6.
- Sousa, J. (2003). Modelling the antigen and cytokine receptors signalling processes and their propagation to lymphocyte population dynamics. Faculdade de Ciências. Lisboa, Universidade de Lisboa.
- Spudich, J. L. and D. E. Koshland, Jr. (1976). "Non-genetic individuality: chance in the single cell." Nature **262**(5568): 467-71.
- Sternberg, P. W. and M. A. Felix (1997). "Evolution of cell lineage." Curr Opin Genet Dev **7**(4): 543-50.
- Suel, G. M., R. P. Kulkarni, et al. (2007). "Tunability and noise dependence in differentiation dynamics." Science **315**(5819): 1716-9.

- Swain, P. S., M. B. Elowitz and E. D. Siggia (2002). "Intrinsic and extrinsic contributions to stochasticity in gene expression." Proc Natl Acad Sci U S A **99**(20): 12795-800.
- Thattai, M. and A. van Oudenaarden (2004). "Stochastic gene expression in fluctuating environments." Genetics **167**(1): 523-30.
- Volfson, D., J. Marciniak, et al. (2006). "Origins of extrinsic variability in eukaryotic gene expression." Nature **439**(7078): 861-4.
- Wolf, D. M., V. V. Vazirani and A. P. Arkin (2005). "Diversity in times of adversity: probabilistic strategies in microbial survival games." J Theor Biol **234**(2): 227-53.
- Wolf, D. M., V. V. Vazirani and A. P. Arkin (2005). "A microbial modified prisoner's dilemma game: how frequency-dependent selection can lead to random phase variation." J Theor Biol **234**(2): 255-62.

