specificity of genomics-based protein-function prediction, although whether specific experimental testing of protein function prediction will ever catch up with the large number of function predictions remains to be seen.

References

1 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
2 Teichmann, S. and Babu, M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* 20, 407–410
3 van Noort, V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.* 19, 238–242
4 Stuart, J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255
5 Kelley, B.P. *et al.* (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11394–11399
6 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736
7 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
8 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
9 Dwight, S.S. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69–72
10 Yu, H. *et al.* (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* 32, 328–337.
11 Peng, W.T. *et al.* (2003) A panoramic view of yeast noncoding RNA processing. *Cell* 113, 919–933
12 Nitta, M. *et al.* (2000) A novel cytoplasmic GTPase XAB1 interacts with DNA repair protein XPA. *Nucleic Acids Res.* 28, 4212–4218
13 Huh, W.K. *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature* 425, 686–691
14 Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41 (http://www.biomedcentral.com/1471-2105/4/41)
15 Ghaemmaghami, S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425, 737–741
16 Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453
17 Troyanskaya, O.G. *et al.* (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U. S. A.* 100, 8348–8353
18 Bozdech, Z. *et al.* (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* 1, E5
19 Le Roch, K.G. *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301, 1503–1508
20 Lasonder, E. *et al.* (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419, 537–542
21 Florens, L. *et al.* (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419, 520–526
22 Coombs, G.H. *et al.* (2001) Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends Parasitol.* 17, 532–537
23 Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12123–12128
24 Pereira-Leal, J.B. *et al.* (2004) Detection of functional modules from protein interaction networks. *Proteins* 54, 49–57
25 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.* 12, 37–46
26 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32 (Database issue), D277–D278

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.06.003

Of statistics and genomes

Diethard Tautz[1] and Michael Lässig[2]

[1]Institut für Genetik der Universität zu Köln, Weyertal 121, 50931 Köln, Germany
[2]Institut für theoretische Physik der Universität zu Köln, Zülpicherstrasse 77, 50937 Köln, Germany

Higher organisms have more genes and larger genomes than simple organisms. This statement sounds almost too trivial to ask the question: why? But there are at least two different answers. Either there is an inherent necessity to increase genome size when more complexity is required or genome size increases because of other reasons that then enable complexity to 'latch on'. Recently, an article by Lynch and Conery, which used arguments of evolutionary population dynamics, proposed that low population size leads to larger genomes. This then provides the opportunity to generate more complex organisms.

The analysis by Lynch and Conery [1] is an excellent example of how important it is to keep the basic predictions of the neutral [2] and nearly neutral theory [3] in mind if one wants to interpret patterns of evolution. These theories describe the statistical fluctuations in finite populations. They emerge as a cornerstone of molecular biology, providing the mathematical framework to place and understand the increasing flood of sequence and genome comparisons. Although the neutral and nearly neutral theories have many complex statistical facets, the main formulae are beautifully simple. They can be viewed as being analogous to formulae in physics that are based on the statistical principles of randomly behaving single units. The general gas theory might serve as an example (Box 1).

Meteorologists use the general gas formula to put the large number of recorded weather data into context, although the prediction of tomorrow's weather is not directly derived from it. Meterologists become more accurate at describing large weather trends by relating large datasets of parallel measurements to each other. Biologists can use the formulae of the (near-) neutral theory to describe large evolutionary trends on the basis of the increasing number of population genetic datasets.

Corresponding author: Diethard Tautz (tautz@uni-koeln.de).
Available online 17 June 2004

Box 1. Gas theory versus neutral theory

In an ideal gas, the movement of each particle becomes random through its frequent collisions with other particles. Therefore, an exact microscopic description of the particles is neither possible nor is it desirable. However, for a macroscopic number of particles, the random fluctuations average out. This results in an exact formula, which relates the variables pressure, volume and temperature to each other:

$$p \times V = n \times R \times T \quad \text{[Eqn I]}$$

p is pressure, V is volume, n is the number of molecules (expressed in mol), R is the general gas constant and T is temperature.

In an ideal population, each individual (chromosome, gene or locus) acts essentially randomly with respect to mutation and reproduction. For a large number of individuals, the statistical fluctuations average out. An exact formula emerges that relates the variables heterozygosity, population size and mutation rate to each other:

$$H = 4 \times N_e \times u/(1 + 4 \times N_e \times u) \quad \text{[Eqn II]}$$

H is heterozygosity, N_e is the effective population size and u is the mutation rate.

Let us draw a few analogies between these two formulae. First, both are macroscopic laws based on a statistical description of the 'microscopic' world. They relate averages. Deviations from these averages can only be neglected in systems that are sufficiently large. (However, the statistical fluctuations in smaller systems can be described as well in both theories.) Second, both theories describe some kind of diffusion. If we know the position of a given particle at some initial time, we can not predict its future movement exactly but we can give a probability distribution. If we know the exact allele composition at a given locus at some initial time, we can not predict its exact fate but we can give a probability distribution for the future composition of alleles. Thus, one can relate the diffusion term in the two formulae: it is proportional to the temperature T for the gas and proportional to $1/N_e$ for a population. Finally, both formulae are based on idealizations. Realistic gases contain interactions between the particles, leading to correlations of their movements. Realistic genes are linked to neighboring genes, and these interactions can be observed through specific linkage disequilibria.

Neutral evolution and scaling

Just as an ideal gas is the minimal model of a many-particle system in physics, the neutral theory is the minimal model for the evolution of a population. Its beauty and strength lies in the small number of parameters: everything depends on the effective population size N_e and the mutation rate u. The near-neutral theory includes selective forces and adds, in the simplest cases, just one more parameter, the selection coefficient s. Does such a simple framework explain at least some facets of the bewildering complexity in biological systems? Going from the ideal gas to ever more intricate many-particle systems, physicists have learned over the past decades that complexity and simple rules are not a contradiction in terms. Surprisingly, simple scaling theories often emerge out of complicated systems, particularly if they are not close to the (thermodynamic) equilibrium of an ideal gas. The (near-)neutral theory is an excellent example of a non-equilibrium scaling theory in biology, and Lynch and Conery have shown how far its consequences might reach.

One of the most important lessons of the neutral theory is that population size matters. The smaller a population is, the easier it is to loose or fix a given allele by chance. The exact probability for the fixation of a neutral allele is $1/2N_e$ in a population of sexually reproducing organisms. Ohta's nearly neutral theory [3] shows that the selection of beneficial alleles can only occur if the selection coefficient is $> 1/2N_e$. Thus, the larger a population is, the more likely it will be able to take advantage of small positive selection coefficients [3,4].

The scaling parameter $N_e u$

Lynch and Conery [1] have compiled data on population size for numerous taxa. This is not as easy a task as it might seem because the genetically effective population size can not be determined by simply counting individuals. For example, the effective population size of a single colony of bacteria is ~ 1, although it can contain 10^{12} individuals, whereas the effective size of the human population with $\sim 6 \times 10^9$ individuals is 10^4–10^5 [5]. In fact, the effective population size is again a parameter that can be inferred from the formulae of the neutral theory. The long-term number of segregating polymorphisms depends only on the product of u and N_e, so it is possible to determine at least the composite parameter $N_e u$ by measuring polymorphisms in a population. It is a general feature of scaling theories that several system variables depend in a simple way on the basic parameters.

Lynch and Conery present several examples of how the $N_e u$ parameter can be correlated to other parameters. First, they show that there is a significant correlation between $N_e u$, genome size and gene number. Multicellular eukaryotic organisms have higher gene numbers and lower population sizes than unicellular prokaryotes. This might seem obvious, but it actually contradicts a *prima facie* expectation of the neutral theory: organisms with a larger population size should be able to retain many genes with specialized or redundant functions (i.e. with small selection coefficient s [3]). But Lynch and Conery argue that the increase in gene number is mainly due to gene duplication and the longer lifetimes of duplicate genes in small N_e organisms. They interpret this as being due to subfunctionalization: the duplicated copies each acquire slightly deleterious mutations, which then require that both copies are retained to produce the full function of the original gene. Because the fixation of slightly deleterious mutations is much less likely in large N_e populations, the probability of complete loss of one of the copies is increased in such populations. Thus, we have a testable hypothesis: in multicellular eukaryotes, gene duplication should normally result in subfunctionalization, an effect that is actually observed frequently [6].

The role of selection

The number and length of introns is another parameter that correlates with $N_e u$. The presence of introns should be slightly disadvantageous to the organism. The extra burden is confined to the probability that one of the nucleotides that is required

for correct splicing is mutated and thus incapacitates the whole gene. This potential selective disadvantage can be calculated. Lynch and Conery suggest that the expected burden of extra introns is too small to enable selection against them in organisms with $N_e u < 0.015$. This is indeed close to the observed crossover scale of intron-rich genomes in eukaryotes. Thus, the expectation of the neutral theory appears to be fulfilled in this case, namely that selection against slightly deleterious mutations is not effective in small populations [3].

Of course, more genes and the presence of introns should have long-term beneficial effects because this enables the generation of increased diversity. Lynch and Conery suggest that these assets of higher eukaryotic organisms might originally have been the neutral consequence of lowered population size in some ancestral species. The inevitable increase in genome size in these ancestral species then enabled them to increase their complexity, whereas those species that retained large populations sizes remained restrained by their evolutionary efficiency of selecting against weakly deleterious mutations. Krakauer and Plotkin [7] have also suggested a model along these lines, in which they suggest that redundancy and antiredundancy mechanisms evolve depending on population size.

Unresolved issues

A few contentious points remain that will need further attention. One is the ongoing debate on possible substructures in populations of prokaryotic organisms [8], which has ramifications for assessing $N_e u$. Another is the role of gene duplication versus orphan gene evolution in determining gene numbers, or the question of how to calculate the mutational load of introns. It appears that some organism groups do not fit the pattern, for example, flagellates and ciliates, which can have large genomes and large population sizes. To put this point in perspective, physicists' experiences with disordered systems can be useful. Random disorder makes it impossible to predict the fate of each single history and creates a spectrum of exceptions but often does not invalidate the scaling of the average and the 'typical'. A more subtle point is Lynch and Conery's basic tenet that among the three fundamental variables N_e, u and s, it is the effective population size that has the dominant effect on genome size. It is difficult to gauge, at this point, the long-term variations of mutation rates or the relevant selection pressures that might themselves depend on the genome size. Eventually a more complex scaling picture might emerge – but one that still has only a few basic parameters.

Lynch and Conery's analysis also highlights the question of regulatory evolution. The proposed subfunctionalization events might have happened at the level of regulation, for example through the degradation of transcription-factor-binding sites. If subfunctionalization were the only process, we should see higher taxa with a lot of genes, but their transcriptional interactions should remain limited. However, the opposite is the case; the complexity of regulation increases in higher taxa, as shown by a proliferation of binding sites in their regulatory DNA [9,10]. To explain this within the framework of the near-neutral theory, one needs to invoke an additional level of complexity, such as the network properties of gene interactions [11].

Concluding remarks

The beauty of Lynch and Conery's article is clearly the consequent application of the mathematical principles of the (nearly) neutral theory to explain the patterns of organismic evolution. Connecting a general theory to real data should be done much more often in biology. The structure and evolution of enhancers and regulatory interactions, the fate of duplicated genes, the evolution of resistance, quantitative traits and the identification of genetic diseases all depend on understanding the patterns that are predicted by the neutral evolutionary theory. This type of statistical theory has proved to be extremely successful in physics. It looks as if it will also become a valuable way of thinking for molecular biologists.

References

1 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
2 Kimura, M. (1986) DNA and the neutral theory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312, 343–354
3 Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286
4 Tautz, D. (2000) A genetic uncertainty problem. *Trends Genet.* 16, 475–477
5 Wall, J.D. (2003) Estimating ancestral population sizes and divergence times. *Genetics* 163, 395–404
6 Massingham, T. *et al.* (2001) Analysing gene function after duplication. *BioEssays* 23, 873–876
7 Krakauer, D.C. and Plotkin, J.B. (2002) Redundancy, antiredundancy, and the robustness of genomes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1405–1409
8 Spratt, B.G. *et al.* (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr. Opin. Microbiol.* 4, 602–606
9 Arnosti, D.N. (2003) Analysis and function of transcriptional regulatory elements: insights from *Drosophila*. *Annu. Rev. Entomol.* 48, 579–602
10 Bolouri, H. and Davidson, E.H. (2002) Modeling transcriptional regulatory networks. *BioEssays* 24, 1118–1129
11 Ohta, T. (2003) Evolution by gene duplication revisited: differentiation of regulatory elements versus proteins. *Genetica* 118, 209–216

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.06.002