24 Yu, K. *et al.* (2003) R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* 4, 442–451
25 Reaban, M.E. and Griffin, J.A. (1990) Induction of RNA-stabilized DNA conformers by transcription of an immunoglobulin switch region. *Nature* 348, 342–344
26 Daniels, G.A. and Lieber, M.R. (1995) RNA:DNA complex formation upon transcription of immunoglobulin switch regions: implications for the mechanism and regulation of class switch recombination. *Nucleic Acids Res.* 23, 5006–5011
27 Shinkura, R. *et al.* (2003) The influence of transcriptional orientation on endogenous switch region function. *Nat. Immunol.* 4, 435–441
28 Longacre, A. and Storb, U. (2000) A novel cytidine deaminase affects antibody diversity. *Cell* 102, 541–544
29 Yoshikawa, K. *et al.* (2002) AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. *Science* 296, 2033–2036
30 Okazaki, I.M. *et al.* (2003) Constitutive expression of AID leads to tumorigenesis. *J. Exp. Med.* 197, 1173–1181
31 Michael, N. *et al.* (2003) The E box motif CAGGTG enhances somatic hypermutation without enhancing transcription. *Immunity* 19, 235–242
32 Ito, S. *et al.* (2004) Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1975–1980
33 McBride, K.M. *et al.* (2004) Somatic hypermutation is limited by CRM1-dependent nuclear export of activation-induced deaminase. *J. Exp. Med.* 199, 1235–1244
34 Brar, S. *et al.* Activation-induced cytidine deaminase (AID) is actively exported out of the nucleus but retained by the induction of DNA breaks. *J. Biol. Chem.* (in press)

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.06.008

Comparative genomics for reliable protein-function prediction from genomic data

Martijn A. Huynen, Berend Snel and Vera van Noort

Nijmegen Center for Molecular Life Sciences, University Medical Center St Radboud and Center for Molecular and Biomolecular Informatics, PO Box 9010, 6500 GL Nijmegen, The Netherlands

Genomic data provide invaluable, yet unreliable information about protein function. However, if the overlap in information among various genomic datasets is taken into account, one observes an increase in the reliability of the protein-function predictions that can be made. Recently published approaches achieved this either by comparing the same type of data from multiple species (horizontal comparative genomics) or by using subtle, Bayesian methods to compare different types of genomic data from a single species (vertical comparative genomics). In this article, we discuss these methods, illustrating horizontal comparative genomics by comparing yeast two-hybrid (Y2H) data from *Saccharomyces cerevisiae* with Y2H data from *Drosophila melanogaster*, and illustrating vertical comparative genomics by comparing RNA expression data with proteomic data from *Plasmodium falciparum*.

Functional genomics data, derived from proteomics and transcriptomics, enable us to have an unprecedented view of global cellular activity. However, these data are 'noisy': they miss many of the true protein interactions and they also report numerous protein interactions that are false. Fortunately, computational analysis of these data can improve our ability to extract reliable predictions from them. Horizontal comparative genomics achieves this by comparing multiple datasets of the same type that are derived from different species. It thus compares not only two independent 'human' experiments but also experiments performed by evolution. It can help answer whether genomic data indicate that proteins from two orthologous groups functionally interact with each other in multiple species.

A classic example of horizontal comparative genomics used the conservation of gene order in prokaryotes as an indication of the co-regulation of proteins in multiple species [1]. This principle has now been applied to gene co-expression data that were determined from array data and to protein–protein interaction data that were obtained through Y2H screens. The likelihood that the observed links between the proteins are biologically meaningful increases dramatically when gene co-expression between orthologous mRNAs is conserved between *Saccharomyces cerevisiae* and *Caernohabditis elegans* [2,3] or among *S. cerevisiae, C. elegans, D. melanogaster* and *Homo sapiens* [4] or when Y2H interactions are conserved between *Helicobacter pylori* and *S. cerevisiae* [5]. This likelihood is measured as the fraction of proteins (among those co-expressed or Y2H interacting proteins whose functions are known) that are part of the same complex, pathway or biological process.

Interestingly, another type of evolutionary conservation, conservation of co-expression or Y2H interaction after parallel gene duplication in one species, leads to a similar increase in the reliability of the predictions that can be made [3,5]. Conservation of co-expression can be used to predict protein function and functional interactions reliably [3,4] and some predictions have been verified experimentally [4].

Function prediction by conserved interaction

There are some conceptual and technical issues that are involved in comparing genomic data from different species,

Corresponding author: Martijn A. Huynen (M.Huynen@cmbi.kun.nl).
Available online 19 June 2004

which we illustrate by comparing the recently published Y2H data from *D. melanogaster* [6] with Y2H data from *S. cerevisiae* [7,8]. Similar to observations made in other horizontal comparative genomics analyses, we found that detecting a Y2H interaction between two orthologous groups in two species dramatically increases the likelihood that they interact functionally (Figure 1); an inspection of the list of conserved interactions indicates that they are all physical interactions. The total number of conserved interactions is however rather low (Figure 1), indicating the high reliability and low sensitivity of using the conservation of Y2H interactions to predict physical interactions between proteins reliably. Furthermore, based on the *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org/) [9], the list of conserved Y2H interactions contains a few proteins for which the biological processes and the molecular functions are unknown (13 proteins, 5% of the proteins in conserved interactions), compared with the complete genome (27% genes with unknown function). This parallels the observation in the *S. cerevisiae* interaction network, where proteins of known function have more interaction partners than those of unknown function [10].

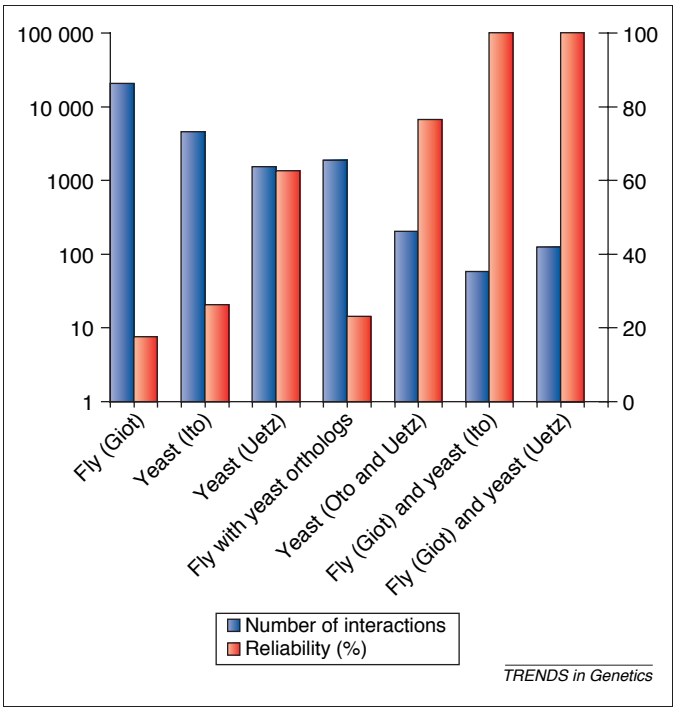

Figure 1. The reliability of separate and combined yeast two-hybrid (Y2H) data for the prediction of functional interactions between proteins (left), and the total number of interactions detected (right). Reliability is measured as the number of proteins for which Y2H interactions are observed (on the same Kyoto encyclopedia of genes and genomes (KEGG) [26] pathway-map) divided by the total number of proteins for which Y2H interactions are observed (in the KEGG database; http://www.genome.ad.jp/kegg). The three columns on the left-hand side contain data from the *Drosophila melanogaster* dataset of Giot *et al.* [6], the *Saccharomyces cerevisiae* datasets from Ito *et al.* [8] and from Uetz *et al.* [7]. In the center are the columns that give the reliability and the number of Y2H interactions that were measured in *D. melanogaster* for which both proteins have orthologs in the *S. cerevisiae* genome. The columns on the right show the reliability for the Y2H interactions that are observed in two datasets. Orthology was defined using the Clusters of Orthologous Groups (COG) database. Conservation of Y2H interaction between *S. cerevisiae* and *D. melanogaster* leads to a greater increase in reliability than the independent observation of that interaction in the two datasets from *S. cerevisiae*, and the fact that two Y2H interacting *D. melanogaster* proteins both have an ortholog in *S. cerevisiae* has little effect on the reliability of the interaction. Note, however, that the reliability of conserved Y2H interactions does come at the price of sensitivity: <200 interactions are conserved between the combined *S. cerevisiae* and *D. melanogaster* datasets. Using the more restrictive best bi-directional hits, the overlap is <100 (Table 1). For complete lists of the conserved interactions and the effects of using different orthology definitions see http://www.cmbi.kun.nl/~huynen/ConservedY2H.

Thus conserved interactions often give credence to cases for which there is already at least some experimental evidence, such as the interaction of the *S. cerevisiae* protein TSR2 and its ortholog in *D. melanogaster* (CG14543) with ribosomal protein 26S in either species. A role for YLR435w in ribosomal maturation would be consistent with the accumulation of 20S rRNA in YLR435w knockouts in *S. cerevisiae* [11] and an interaction with the ribosomal protein 26S in *S. cerevisiae* has also been observed using tandem affinity purification (TAP)-tagging [11].

Nevertheless, using overlapping datasets, one can make new, reliable protein–protein interaction predictions. One example of a 'new' interaction is between the xeroderma pigmentosum group A binding GTPase (XAB1) (CG3704 in *D. melanogaster*) and the hypothetical protein YOR262w/CG10222, which also contains a GTPase domain. XAB1 has been observed to interact with the DNA-repair protein XPA1 and is thought to be required for its import into the nucleus [12], suggesting a function in nuclear import for YOR262w/CG10222. The two proteins share other, albeit weak, genomic links: in *S. cerevisiae*, both proteins are cytoplasmatic [13] and are essential for the cell [9]. Furthermore, they have the same phylogenetic distribution, possessing orthologs in all eukaryotic genomes sequenced to date.

More than just more data

Is combining data from different species really more than just combining multiple independently generated datasets to filter out the experimental noise? There are some indications that this is the case: Stuart and coworkers showed that omitting parts of the co-expression data in each species did not significantly affect their ability to reliably predict the proteins that were part of the same pathway [4]. By contrast, omitting one or more species from the data drastically reduced this predictive value [4]. Furthermore, although taking the overlap into account between two different Y2H datasets for *S. cerevisiae* also leads to an increase in the likelihood of detecting a real interaction, it does not match the level that is obtained by comparing datasets from multiple species (Figure 1). Note however that when two genes are co-expressed in *S. cerevisiae* the fact that they both have orthologs in *C. elegans* in itself already leads to a higher likelihood of interaction, although not as high a likelihood as when these orthologs are also co-expressed in *C. elegans* [3]. Thus, the (same) widespread, phylogenetic distribution of genes appears to be responsible in part for the increase in the predictive value of finding conserved co-expression between them. This effect is however not observed for the Y2H data (Figure 1), where the presence of orthologs in the other species does not significantly affect the reliability of the interaction.

Quantifying the amount of conservation and determining orthology

The amount of conservation of co-expression between *S. cerevisiae* and *C. elegans* is low ($<10\%$) [2,3], albeit

Table 1. The overlap among the number of yeast two-hybrid interactions between two *Saccharomyces cerevisiae* datasets and one *Drosophila melanogaster* dataset[a,b]

Dataset comparison	Protein interactions (both proteins present in the other dataset)	Conserved interactions	Percentage of conserved interactions	Mean conserved interactions
Ito versus Uetz	858; 697	201	23.4%; 28.8%	26.1%
Ito versus Giot	229; 394	45	19.6%; 11.4%	15.5%
Uetz versus Giot	120; 168	33	27.5%; 19.6%	23.5%

[a]In calculating the percentage of overlap between datasets only interactions for proteins that appeared in both datasets were taken into account (i.e. the number of interactions that was observed in both *S. cerevisiae* datasets was divided by the number of interactions in the Ito set, for which both proteins were present, although not necessarily interacting with each other, in the Uetz set). Orthology relations between the *S. cerevisiae* and the *D. melanogaster* were determined by best bidirectional hits among the homologous relationships ($E < 0.01$, local sequence alignment). The percentage overlap between the *S. cerevisiae* datasets and the *D. melanogaster* dataset (24% and 16%) are close to those between the *S. cerevisiae* datasets themselves (26%). This suggests that the low levels (24% and 16%) of conservation of physical interaction between *S. cerevisiae* and *D. melanogaster* can, to a large extent, be attributed to the low reproducibility of yeast two-hybrid (Y2H) interactions in general; therefore, physical interaction between proteins is highly conserved in evolution.
[b]For more information on the datasets used, see Refs [6–8].

significant [3]. It is not clear to what extent the small overlap is a reflection of the noisy nature of the data or a true indication of the low conservation of co-regulation. Determining the conservation of protein–protein interactions depends, of course, on how orthologous relationships between proteins are determined. Using a restrictive measure of orthology, such as best bidirectional hits, there are 33 and 45 conserved interactions between the *D. melanogaster* and the Uetz dataset [7] and between *D. melanogaster* and Ito [8] dataset, respectively. When dividing by the number of Y2H-interacting *D. melanogaster* proteins whose orthologs are actually present in the *S. cerevisiae,* Uetz and Ito Y2H datasets, there are 24% conserved interactions between *D. melanogaster* and the Uetz dataset and 16% between *D. melanogaster* and the Ito dataset (Table 1). These percentages of conserved interactions are substantial when compared with the 26% of interactions that are 'conserved' between the Y2H *S. cerevisiae* datasets. Best bidirectional hits do not necessarily identify functionally equivalent orthologs, especially in the case of gene duplication and varying rates of evolution, and more inclusive measures might identify extra conservation. Using the more inclusive eukaryotic orthologous groups (KOG) [see Clusters of Orthologous Groups (COG) http://www.ncbi.nlm.nih.gov/COG] [14] to define orthology relationships, the number of 'conserved' interactions between *D. melanogaster* and *S. cerevisiae* increases from 33 to 39 (Uetz dataset) and from 45 to 51 (Ito dataset) (see http://www.cmbi.kun.nl/~huynen/ConservedY2H). Therefore, the issue of how to determine orthology relationships between genomes is becoming less academic because similar functional genomic data for multiple species are available, and we can finally compare the various orthology algorithms for their sensitivity and selectivity.

Vertical genomics – different data from one species

By only analyzing the overlap between genomic data from different species we, of course, ignore biologically relevant, species-specific interactions. To detect such interactions reliably, one can combine different types of genomic data from one species. In doing so one faces several challenges. First, the predictive values of various types of genomic data vary widely not only among different types of genomic data but also within one set of genomic data (e.g. in co-localization data, where the predictive value for protein interaction depends strongly on where in the cell the proteins co-localize [13]). Second, datasets tend to be incomplete: they tend to cover only a fraction of the genes [10] (i.e. except for RNA-expression data or genome data). Finally, there are intrinsic correlations between the data, for example, between expression data on the RNA level and on the protein level [15].

Recently published approaches tackled these challenges by combining the genomic data in a Bayesian framework [16,17]. A Bayesian approach uses a set of known interactions and known non-interactions (e.g. the proteins are in different cell compartments) to estimate how best to combine the various types of data, instead of just 'blindly' combining them as was illustrated previously for the Y2H data. Furthermore, the quality of the predictions is expressed as the likelihood that two proteins interact relative to two randomly chosen proteins and is not an absolute probability as shown in Figure 1. When the genomic data are, in principle, independent (e.g. localization and expression data) they are combined in a so-called Naïve Bayesian approach, in which the likelihood that two proteins interact for the separate datasets are multiplied by each other to obtain a combined likelihood. A full Bayesian analysis does not assume the independence of the data and estimates the likelihood by directly comparing combinations of various (binned) values of the data with a set of known interactions and known non-interactions.

A Bayesian network for *Plasmodium falciparum*

We illustrate the Bayesian network approach with an analysis of genomic data for *P. falciparum* for which two gene-expression datasets [18,19] and two proteomics datasets [20,21] have been published. Because these data all reflect gene expression, either measured directly as transcript or indirectly as protein, and are therefore not independent, they have to be combined in a full Bayesian framework: the likelihood of protein interaction has to be directly estimated for combinations of correlations between the genes in the separate datasets. Combining the data in this manner leads to an increase in the likelihood of the predictions that can be made in two ways (Figure 2): (i) protein interactions that are supported consistently by all datasets are more probable than those that are only supported by one dataset; and (ii) with an

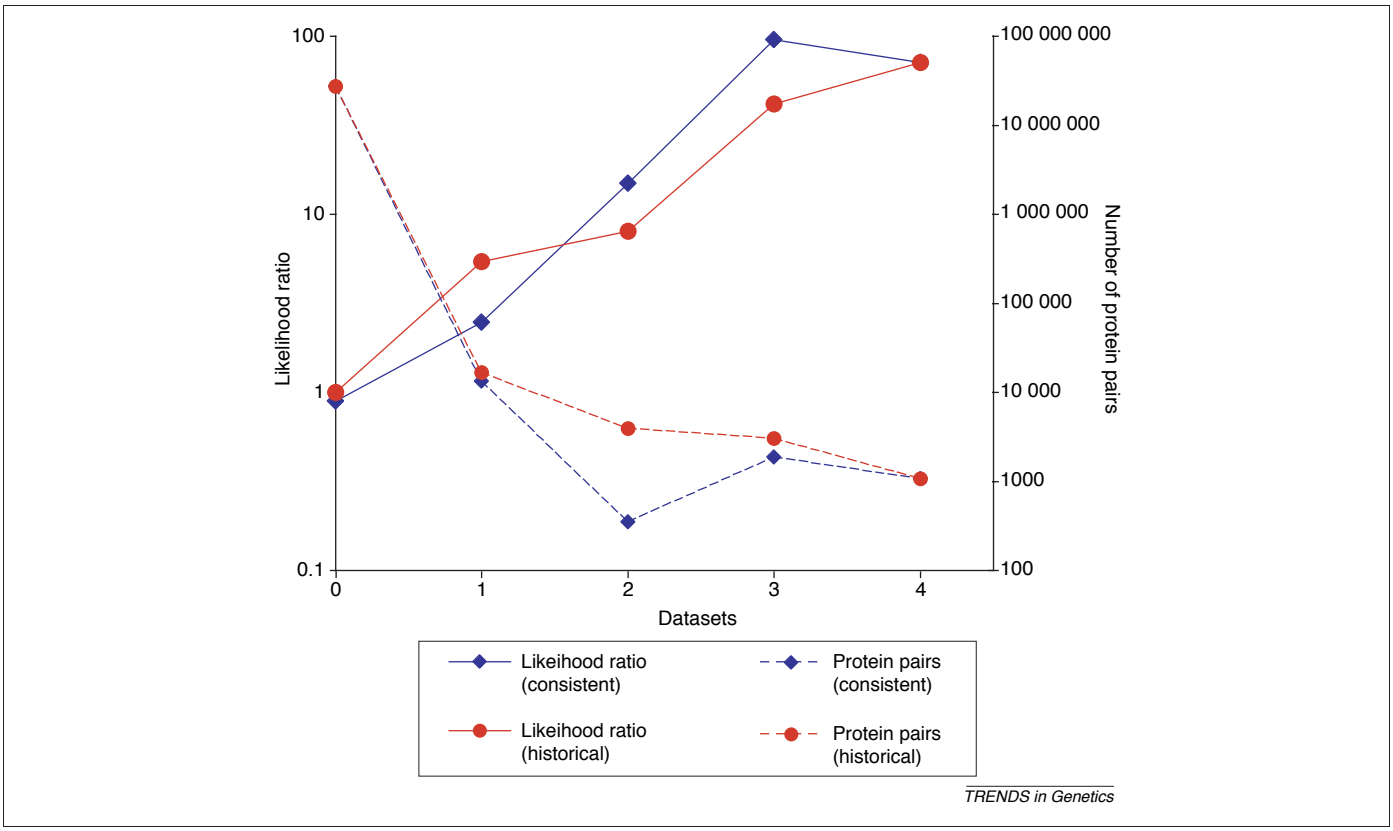

Figure 2. The increase in the likelihood value for the interaction of *Plasmodium falciparum* proteins observed by consistently correlated expression over four datasets (unbroken lines) and the number of predicted protein pairs at that likelihood value (broken lines). A Bayesian analysis for the two protein and two RNA expression datasets for *P. falciparum* was performed by calculating pair-wise Pearson-correlation coefficients for all genes in each dataset separately. For the proteomic data, these are based on the number of different peptides detected per protein [21] (non-tryptic peptides excluded). Next pairs of genes for each set were divided into two classes: high correlation and low correlation. Finally, for combinations of 'high correlation' and 'low correlation', the relative likelihoods that the proteins interact were estimated based on co-occurrence in Kyoto encyclopedia of genes and genomes (KEGG) maps (http://www.genome.ad.jp/kegg). The 'consistent' line (blue) shows the likelihood that proteins interact based on a high correlation in only one dataset (and a low correlation in the other three), a high correlation in two datasets (and a low correlation in the other three) and so on. The red line is a historical reconstruction of how the addition of datasets has increased the possibility to predict protein interaction with high likelihood. It is calculated for a high correlation in the first dataset (irrespective of the other three), in the first two (irrespective of the other two) and so on. The order of adding the datasets, from left to right is: proteomics from Lasonder *et al.* [20] and from Florens *et al.* [21], RNA expression from Le Roch *et al.* [19] and from Bozdech *et al.* [18], with correlation thresholds between 'high correlation' and 'low correlation' set at 0.8, 0.75, 0.75 and 0.8, respectively. There are not enough data to divide the correlations into more categories than high and low. For a complete listing of interaction likelihoods of protein pairs, see http://www.cmbi.kun.nl/~huynen/PlasmodiumData. The non-independence of the data is reflected in the saturation of the curve, independent data would, in principle, produce a straight line with the increase in likelihood. The results illustrate the value of having more data for the prediction of protein–protein interactions in *P. falciparum*, even when those are correlated.

increase in data over time the quality of the predictions improves (i.e. having more data enables us to making more likely predictions).

Function prediction in *Plasmodium falciparum*

P. falciparum protein functions can be predicted more reliably using the integrated data. A typical example is PFI0895c, a protein that is homologous to subunit 5 of translation elongation factor 3 (elF-3 epsilon), which interacts with the ribosome, and is homologous to subunit RPN8 of the 26S proteasome regulatory complex. Both at the RNA and at the protein level, PFI0895c shows a correlated expression with ribosomal proteins L27, L21e and Sa. An annotation of PFI0895c as elF-3 epsilon appears therefore most likely. Potentially more interesting are predictions for proteins that are specific to the *Plasmodium* genus such as PFI0555c, which is expressed with two proteins that are involved in protein degradation – the aspartic proteinase and drug target [22] PF14_0075 (plasmepsin IV) and the ornithine aminotransferase MAL6P1.91 – suggesting an additional role for PFI0555c in protein degradation.

Outlook

Comparative genomics is a powerful tool to extract reliable predictions from genomic data. To obtain predictions that are amenable to experimental testing, predictions need not only to be reliable but also more specific than 'protein A is involved in process B' or 'protein A interacts with protein C'. One source of information to make predictions more specific is the topology of the predicted interaction networks. Locally densely connected networks reflect stable physical complexes, whereas less connected networks correlate with signaling pathways and transient interactions [23,24]. With the avalanche of genomic data, instead of merely determining the overlaps, scientists will be in a position to determine the differences and extract biological meaning from those differences. In contrast to stable complexes, transiently interacting proteins appear not to show correlated expression [25]. Alternatively, one might be able to find metabolic pathways in consistently co-expressed but not physically interacting proteins. We will undoubtedly see that more creative combinations of genomic data will increase the

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