The Pharmacogenomics Journal (2001) ●, ●●●–●●●

www.nature.com/tpj

REVIEW

Global perspectives on proteins: comparing genomes in terms of folds, pathways and beyond

R Das
J Junker
D Greenbaum
MB Gerstein

Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA

Correspondence:
MB Gerstein, Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA
Tel: +1 203 432 6105
Fax: +1 203 432 5175
E-mail: Mark.Gerstein@yale.edu

Received: ●● Month 2001
Revised: ●● Month 2001
Accepted: ●● Month 2001

ABSTRACT

The sequencing of complete genomes provides us with a global view of all the proteins in an organism. Proteomic analysis can be done on a purely sequence-based level, with a focus on finding homologues and grouping them into families and clusters of orthologs. However, incorporating protein structure into this analysis provides valuable simplification; it allows one to collect together very distantly related sequences, thus condensing the proteome into a minimal number of 'parts.' We describe issues related to surveying proteomes in terms of structural parts, including methods for fold assignment and formats for comparisons (eg top-10 lists and whole-genome trees), and show how biases in the databases and in sampling can affect these surveys. We illustrate our main points through a case study on the unique protein properties evident in many thermophile genomes (eg more salt bridges). Finally, we discuss metabolic pathways as an even greater simplification of genomes. In comparison to folds these allow the organization of many more genes into coherent systems, yet can nevertheless be understood in many of the same terms. *The Pharmacogenomics Journal* (2001) ●, ●●●–●●●.

Keywords: genome comparison; protein structure; folds; thermal stability

INTRODUCTION

With the advent of new DNA sequencing technology there are as many as 800 organisms for which genomes have been neither completely sequenced or sequencing is in progress. The attention, both public and scientific, has catalyzed a tremendous effort to analyze and compare those genomes that are publicly available.[1–3] This interest is reflected in the large number of genome comparison articles over the last decade. The increase in the number of publications comparing genomes (from 75 in 1990 to 220 ten years later) shows a strong upward trend, suggesting much more of this activity in the future (see Figure 1). The accumulation of sequence data has resulted in a paradigm shift in the biological method; the bottleneck now occurs in data analysis rather than data generation.[4] The analysis of these data will allow researchers to raise, and attempt to answer, many complex biological questions that were not possible to address in the pre-genomic era. This review attempts to briefly outline some rudimentary comparison methods for genome analysis, as well as present some more novel and efficient options for comparing genomes.

TYPES OF GENOME COMPARISON

Comparison Based on Single Sequences

Deciphering a genome is akin to trying to understand a dead language without the help of a Rosetta stone. Fortunately, we are not working from a true *tabula rasa* as biologists have imported tools and methods from other data-heavy sciences. Tools such as Bayesian networks, Self-organizing maps and Hidden Markov Models have allowed for a better understanding of the

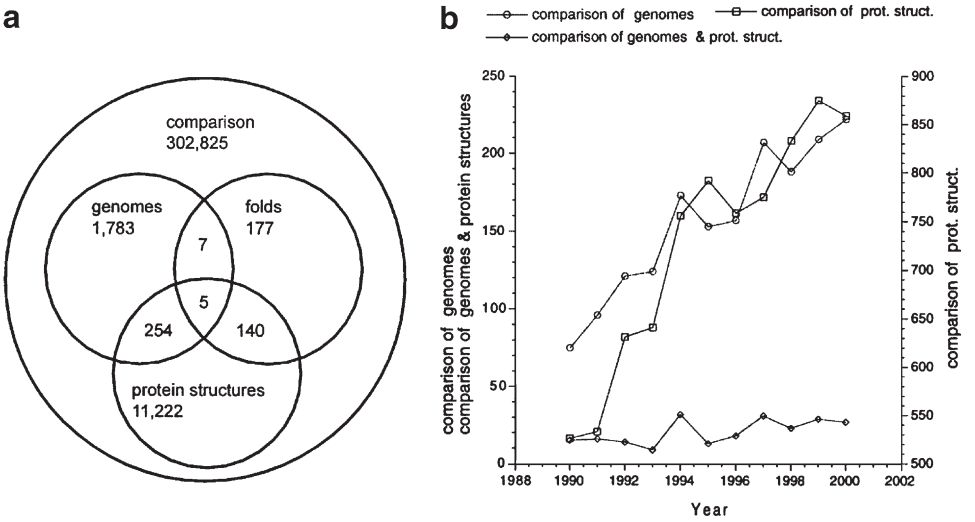

Figure 1 Advantages of organizing sequences into folds. Folds can group a large number of sequences into a smaller number of folds. For instance, there are about 30 000 genes in human, and they can be organized into 1000-fold families. Furthermore, folds can group evolutionarily related sequences.

underlying data. These methods can be used to compare genomes in multiple varied fashions.

Initially researchers used straightforward approaches to compare genomes directly in terms of sequence. These methods searched for: (i) homologues, motifs (eg regulatory or DNA binding) and common oligonucleotide and oligopeptide words;[5–8] (ii) orthologs (see for instance the COGS database;[9,10] (iii) gene duplications;[11–19] and (iv) the occurrence of conserved families in several different genomes.[9,20–24] Several semi- and fully-automated methods have also been developed for comparing whole genome sequences against multiple databases.[25–33]

COMPARISON BASED ON GROUPING SEQUENCES INTO FOLDS

Why Folds?

Efficient genome analysis requires the organization of an enormous number of protein sequences in a systematic and orderly fashion. The most general way of organizing genomes involves clustering the sequences into protein families based on sequence similarity. However, in many instances, sequences, although evolutionarily related, diverge so much that no appreciable homology can be found to group them into the same family. In contrast to groupings based purely on sequence similarity, folds provide for greater simplification in organizing the large amount of genomic data (Figure 2). Furthermore, in many cases, folds define function, and can maintain their function even with mutations in the sequence. Thus, two seemingly divergent sequences, can code for the same fold and, as such, can be grouped together independent of their minimal sequence homology.[34,35]

Genome comparison based on protein structure is important for multiple reasons. First, one can define a structural module precisely, and there is a limited number of motifs as opposed to sequences.[13,36–51] Moreover, analysis of structure can reveal more about distant evolutionary relationships than sequence comparison alone, as structure is more conserved than sequence or function.[52,53] Furthermore, the relationship between sequence similarity and structural similarity is better defined than the corresponding relationship between sequence and function. Finally, an emphasis on structure will help further our knowledge in drug design and molecular disease. The difficulty in identifying drug targets from raw genomic sequence alone is reflected in the low (10%) percentage of pharmaceuticals that are developed through genomic efforts.[54,55] Structural proteomics' computational methods for structure study can overcome some of the limitations of other high throughput experimental methodologies (ie the difficulty in studying proteins due to insolubility or unstable folding.[56] As there is a large degree of structural, and thus functional, homology between completely different sequences, there is obviously a large number of unknown homologies that pharmaceuticals can take advantage of by determining structural and functional features for previously un-annotated proteins.[57] Structures may also help us interpret Single Nucleotide Polymorphisms (SNPs) in coding regions. In particular, they will allow us to make inferences regarding selection, mutation and function of SNPs by comparing similar structures with a range of underlying sequences.

Types of Structural Comparison

Structural comparison can be made on multiple levels. The concept of structure extends from alpha helices and sheets to complex multi-domain motifs to whole proteins and complexes. A more complex structure will be more evolutionarily conserved and will also be more informative in terms of function.

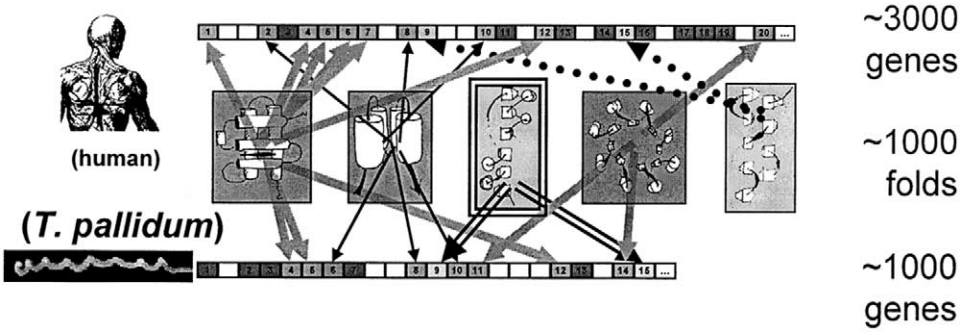

Figure 2 Trend in published research articles on structural genomics. (a) Results of PubMed searches for the keywords 'comparison', 'protein structure', 'genomes' and their combinations. Whereas it is obvious that the number of references to the word 'comparison' (316 824) will be large, the number of publications comparing genomes (2059) or protein structures (11 621) is surprisingly small. The results are illustrated here as subsets. (b) The analysis of the numbers of publications per year regarding comparison of protein structures, genomes and protein structures and genomes reveals that the number is continuously increasing. Especially the number of publications regarding comparisons of genomes has tripled over the past 10 years.

Fold Libraries

A common objective of most of the structural studies is to achieve an understanding of large proteomes in terms of a limited repertoire of structures culled from fold libraries. Manual as well as automatic methods are used for structural alignments as sources for fold databases such as SCOP, FSSP, CATH and HOMALDB.[58–62] Pfam, which catalogs multiple sequence alignments of protein domains or conserved protein regions, is another example of a database used for comparative studies.[63] Pfam is especially useful for automatic detection of remote homology by building profiles via Hidden Markov Models.

Fold Recognition: Comparing Folds to Genomes with Templates

Currently, the PDB can be clustered into [11]360 representative domains. Using structure comparison, one can further cluster the data into 564 folds, giving about two sequence families per fold.[64] Sequence templates, authoritative sequences for a given fold, can be extracted from these fold libraries and used to search the genomes. These templates are used specifically as seeds to build up large sequence alignments from the major databases using standard pair-wise searching tools—eg the popular BLAST and FASTA programs on the Swissprot and GenBank databases.[65–69] A number of methods of transitive comparisons are expected to improve the sensitivity of these pair-wise searches.[65,70,71] Since many of these alignments contain quite a few sequences, they can be fused into a consensus pattern or template using various probabilistic approaches including Hidden Markov Models.[72–77]

PSI-BLAST, in addition to other methods, is used to compare these templates directly against the genomes to find other similar folds and to detect remote homologies.[45,65,78–82] If one finds a close homology, one can obviously use this to model the target protein based on the template information.[83–87]

Approaches to Large-scale Surveys: Common Folds, Shared Folds and Horizontally Transferred Folds

There are many large-scale surveys and comparisons based on folds that have been performed using the above methods. These fold comparisons have provided an important perspective of a finite 'parts list' for different organisms.[88,89] It is argued that with few exceptions, the tertiary structures of proteins adopt one of a limited repertoire of folds.[90–92] As the number of different fold families is considerably smaller than the number of gene families, categorising proteins by fold provides a substantial simplification of the contents of a genome. One can expect that this notion of a finite parts list will become increasingly common in the future genomic analyses.

There are many ways in which genomes have been studied and compared in terms of protein folds (eg, in terms of the most abundant folds). Such 'inventory statistics' can be very useful in understanding the individual characteristics of genomes, particularly of microbial physiology. Similarly, if the results are compared among the organisms, one can obtain knowledge regarding shared folds among those genomes. Similar analyses have been performed to look into such distributions in a number of the recently sequenced genomes.[93,94] As shown in Figure 3, the analysis can be conceptualized in terms of a Venn diagram, similar to those used for studying the occurrence of motifs and sequence families.[46,95] Out of the known folds (564) about half are contained in at least one of the three genomes studied, and 200[93] folds are shared amongst all three genomes. These shared folds presumably represent an ancient set of molecular parts. Protein folds in the worm genome have also been surveyed, revealing that there are about 32 matches per fold and involving a quarter of the total worm ORFs.[96,97] Comparison with other model organisms also showed that the worm is phylogenetically closer to yeast than *E. coli*.[96] Folds were also assigned to the proteins encoded by the genome of *Mycoplasma genitalium*.[98] Studies have been performed to

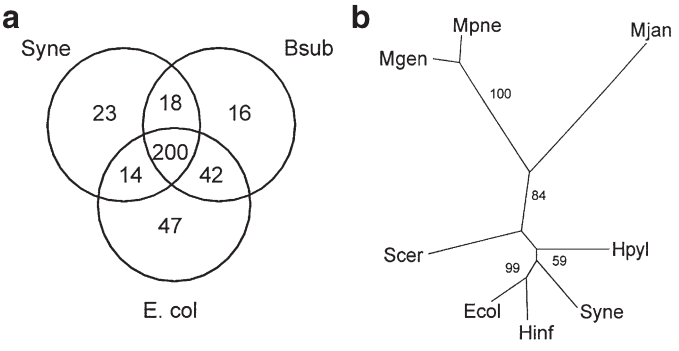

Figure 3 Various ways of comparing genomes in terms of structures. (a) Genomic tree based on the overall occurrence of folds in the genomes, generated by a distance-based method. For each of the microbial organisms the presence or absence of folds was marked with 1 or 0, respectively. (b) Distribution of known folds amongst the genomes. This figure is adapted from Gerstein *et al*.[96] There are almost ~500 known folds, of which almost half of them are shared between all three.

relate these folds with functions.[99] Furthermore, three-dimensional protein folds were also assigned to all ORFs in the recently sequenced genomes hyperthermophilic archaeon and *Pyrobaculum aerophilum*.[100] Efforts have been further made to assign folds for proteins with unknown functions in three microbial genomes *Mycoplasma genitalium, Haemophilus influenzae*, and *Methanococcus jannaschii*.[101] In addition to fold assignment, studies have also addressed the pattern of fold usage across genomes.[93] The sharing of folds across many different genomes can be used to group organisms into cluster trees.[94] These whole-genome trees have a remarkable amount of similarity to the traditional ribosomal tree, despite being based on completely different metrics of similarity (see Figure 3b).

PEDANT and GeneQuiz are two web sites that compile these data automatically.[27,102] Such comparison provides a global view of fold abundance across the organisms and their evolution. Moreover, this comparison can tell us if certain genes had been horizontally transferred between two evolutionarily distant organisms.

This idea of fold comparison is not limited to ORFs, but can also be extended to pseudogenes, ie those genes that are not expressed. In a recent survey of the estimated pseudogene population in the worm genome, the distribution of top protein folds in the proteome and in the predicted pseudogenes showed some notable differences, with a number of folds, in particular that of DNAase I, being much more common in pseudogenes than in expressed genes.[103]

Comparison of Predicted Structure

It is obvious that we can't assign a fold to all expressed sequences in a genome thus limiting any genome comparison based solely on folds. As such there are efforts being made to predict the structure of unknown proteins.[104,105] In addition to homology modeling, there are other prediction methods that have been developed to gain structural information for the sequences that do not have any similarity with a known fold. Unfortunately though, 3D structure prediction based on an '*ab initio*' method has not been very successful.[106–111]

Structure prediction has been most successful with one-dimensional prediction for secondary structure, assigning individual residues in the protein sequence to discrete states like strand, coil or helix. Methods such as GOR (Garnier–Robson–Osguthorpe Secondary Structure Prediction) incorporate multiple sequence information.[112–114] The DSC method (Discrimination of Secondary structure Class) and the method developed by Livingstone and Barton are other popular methods, and tend to give more accurate results.[115,116] Using these predicted secondary structures, multiple genomes have been successfully compared. It was found that genomes have a similar secondary structure composition even through they have different amino acid compositions.[15,88,117]

In addition to predicting helixes and beta sheets, several prediction methods have been developed for transmembrane helices. Some of them are based on parameters derived from the intrinsic properties of amino acids, usually their oil–water transfer energies. A widely used example is the GES hydrophobicity scale.[118] Other authors using different scales, eg the Kyte–Doolittle or the Eisenberg scales, also developed similar prediction methods.[119–123]

A Case Study in Structural Genomics Comparisons: Finding the Unique Featrues of Proteins in Thermophiles

To illustrate how genome analyses can be used to understand the structural properties of proteins, we describe a case study comparing the genome sequences of thermophiles to those of mesophiles.[124] We focus on the question of what are the unique properties of proteins that are stable at high temperature and use this to illustrate various comparative methodologies.

Thermophiles (archaea and a few eubacteria), thrive in high temperatures. It is not well understood how thermophiles stabilize proteins at these elevated temperatures that otherwise denature normal-temperature (10–45°C) mesophilic proteins. Crystallographic studies, as well as structural information obtained through homology modeling, revealed a strong correlation between the number of salt bridges and protein thermal stability.[125–140] There are several ways in which salt bridges can stabilize proteins. Ion pair networks, helix stabilizing salt bridges, salt bridges buried in a hydrophobic core and surface salt bridges between two subunits are among the most frequently encountered types.[126,129,131,135,141,142] Most of these past studies, however, were anecdotal in nature in that they focused on one specific protein rather than a comprehensive population sample. Consequently, it is interesting to see how a comparative genomic analysis could bring a global perspective to such understanding.

The purpose of such a comparison is to find an overall statistical difference for proteins in thermophile genomes in comparison to mesophiles. This global view does not limit the researcher to the evaluation of an isolated individual difference in a particular protein, but rather focuses on overall

differences over the whole genome. The most obvious parameters one can look at are the sequence composition and length of all the ORFs in each genome. Figure 4 shows a simple illustrative comparison of five thermophilic genomes with seven mesophilic genomes in terms of amino acid content. On a primary sequence level, we see that thermophile genomes overwhelmingly have more charged residues than mesophiles. This result becomes more striking when we take into account secondary structure considerations, through prediction of the secondary structure for all ORFs in the genome using standard approaches such as the GOR program. It is generally known that charged residues are associated with stabilizing salt bridges. A further investigation into the secondary aspects of these proteins shows that not only are there more charged residues in general, but this trend is also evident in predicted helices and that the spacing of the charged residues in these helices has a preferred 1–4 arrangement. This 1–4 arrangement is usually associated with intrahelical salt bridges[143,144] (see Figure 5a). To demonstrate the preferred 1–4 arrangement, one can compute a LOD value (ie the odds of having charged residues in a particular spacing relative to a random expectation). These LOD values point to the high probability of salt-bridges in thermophiles compared to mesophiles Moreover, the frequency of salt bridges correlates with the physiological temperature of the organisms such that the number of salt bridges increases with the increase in physiological temperature as shown in Figure 5b. Thus the additional information of secondary structure provides us with a clearer view of how primary sequence differences can be explained as functional differences.

Biases and Sampling

General Issue of Bias in the Databanks

One imperative concern in all large-scale surveys, such as the above protein thermostability example, is that of biases. There are many ways in which a bias can arise in a dataset. One large source of bias is the consequence of investigator preferences, resulting in the over or under representation of certain sequences and structure (eg compare human and fly globins in the GenBank repository). By focusing only on organisms for which complete genomes are known, one can attempt to eliminate this form of bias. However, this will not remedy the biases resulting from sequence repeats. The repetitive charged sequences in the set of thermophilic proteins from the above example could skew those results. Moreover, protein sequences enriched in salt bridges, unique to the thermophile, could be duplicated in the thermophile genomes forming large paralogous families and influencing the results. A similar situation may arise involving only the sequences unique to mesophiles.

In addition to biases in sequence databases, there are also biases in the structural databanks. The selection of proteins in the PDB is biased by the preferences of individual investigators and by the physical constraints imposed by crystallography and NMR spectroscopy. Structures in the PDB are also biased towards certain commonly studied organisms. Another important issue related to bias in the structure databank is that the absolute counts found in a given genome survey are contingent on the evolving contents of the databank. Thus, over time, as more structures are added to the databank, one should expect the basic inventory statistics (eg the most common folds or the number of shared folds) to change.

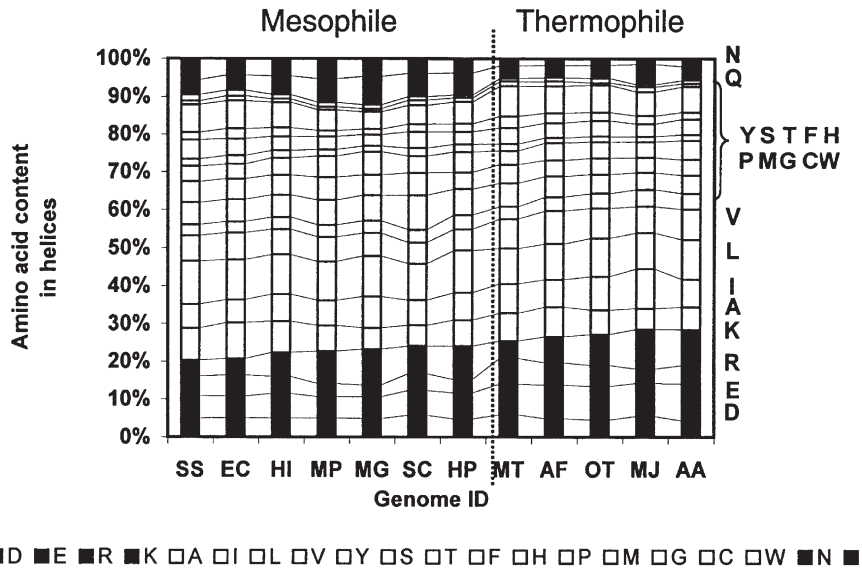

Figure 4 Amino acid composition in helices. Figure is adapted from Das and Gerstein.[124] The blackened area in the figure represents the portion of charged residues E, D, K and R in a helix. This area increases from mesophiles to thermophiles.

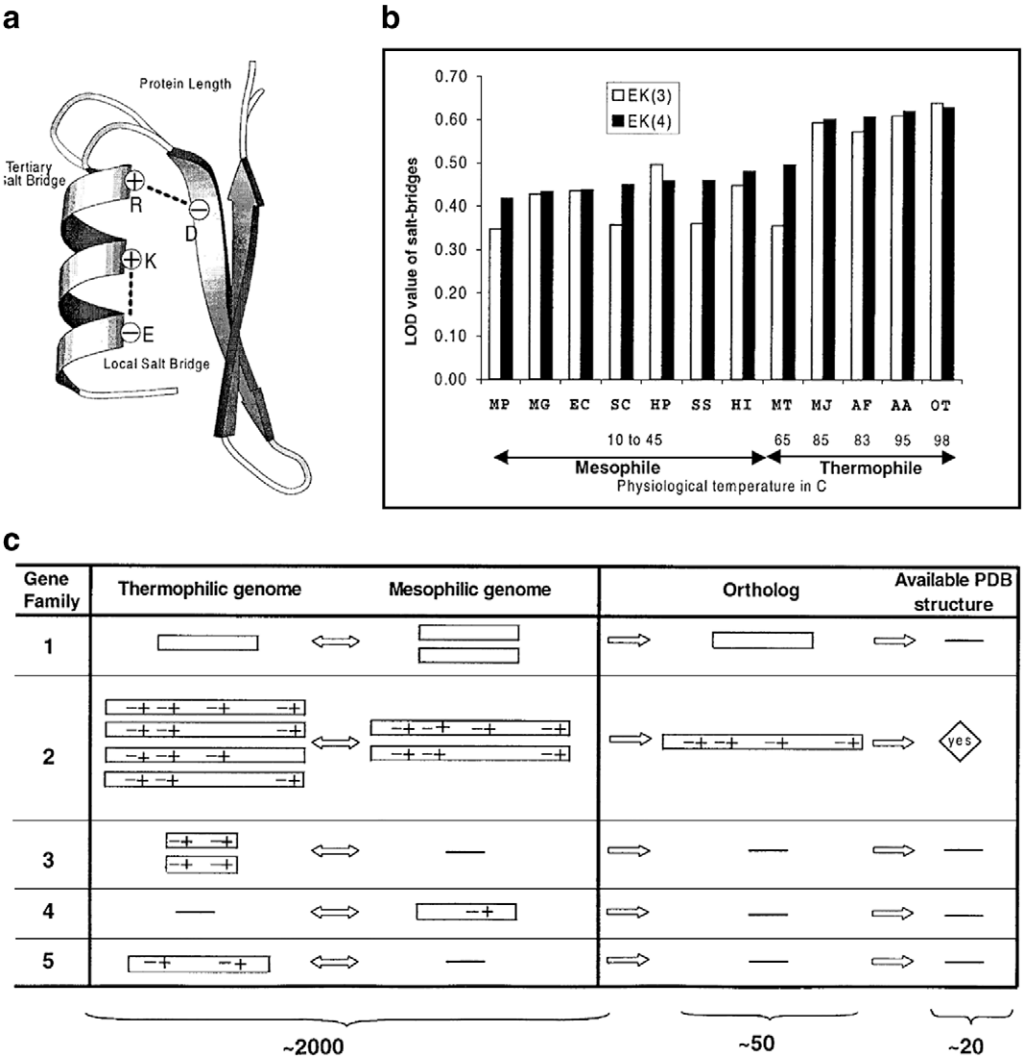

Figure 5 A case study of comparative genome analyses focusing on protein thermostability. Figure is adapted from Das and Gerstein.[124] (a) Intra-helical as well as tertiary salt bridges stabilize protein structure. (b) LOD values increase with the increase in physiological temperatures shown along the horizontal axis. For mesophiles, they are indicated by a range from 10 to 45°C. The two letter codes represent individual genomes: *Pyrococcus horikoshii* (OT3), *Aquifex aeolicus* (AA), *Methanococcus janaschii* (MJ), *Archaeoglobus fulgidus* (AF), *Methanobacterium thermoautotrophicum* (MT), *Haemophilus influenzae* (HI), *Mycoplasma genitalium* (MG), *Mycoplasma pneumoniae* (MP), Helicobacter *pylori* (HP), *Escherichia coli* (EC), *Synechocystis* sp (SS), *Saccharomyces cerevisiae* (SC). (c) This diagram illustrates the strategy of stratified resampling, a method that can be used to eliminate biases. In this salt-bridge example, 52 orthologous proteins were selected (from an assumed size of ~2000) by this method. That is, only those corresponding proteins, which are present in all 12 genomes, were selected, and then only a single representative was actually counted. The dashed lines (—) in the figure show the sequences that are missing for any orthologous group and are thus discarded from calculation. Using this procedure, one can filter out the effect of paralogous sequences as well as sequence repeats that may bias results.

Biases in the Prediction Programs

Cobbling together an 'inventory census' through the use of a disparate collection of tools and patterns creates another type of bias, that of devising consistent scores and thresholds. This is particularly acute in the case of manually derived sequence patterns and motifs, since an expert on a particular fold or motif would expect their pattern to find relatively more homologues than a pattern not constructed by an expert. Applying the same single-sequence procedure to each fold circumvents these problems to some degree. Furthermore, this simplification has the added advantage in that it can be performed automatically without manual intervention and, consequently, can easily be scaled up to deal with much larger datasets.

In addition to biases discussed above, there are also biases integrated into each of the tools used in large-scale analyses. Secondary structure prediction using GOR is statistically based, so that the prediction for a particular residue to be in a given state, say Valine in a helix, is directly based on the frequency that this residue occurs in this state in a data-

base of solved structures (taking into account neighbors at ±1, ±2, and so forth). Therefore, a bias in the sequences in the structure database would be propagated in the structure prediction. The GOR method only uses single sequence information and thus, achieves lower accuracy (65%) than the current 'state-of-the-art' methods (71%) that incorporate multiple sequence information.[3,59,145,146] Moreover, it is not possible to obtain multiple sequence alignments for most of the proteins in the genomes. Consequently, bulk predictions of all the proteins in a genome based on multiple-alignment approaches are skewed, in the same sense as discussed above for multiple-sequence based fold-recognition methods.

How to Deal with Biases in Comparative Study?

In doing genome-wide surveys, one has to be careful to assess the degree to which one's calculated statistics could be biased. Results should be tested and significance should not be assigned without statistical controls and alternate procedures.

Random Resampling

Random sampling procedures can be used to test results to see if they are biased by sequence repeats. By comparing the statistics from randomly selected sequences with the overall results, one can estimate the extent of bias in the database. In the above case study, simulated thermophilic and mesophilic genomes could be made up by randomly drawing protein sequences from two large pools of thermophilic and mesophilic sequences. LOD values obtained from these simulated genomes would reflect the effect of biases. In this specific case no such bias was found.

Stratified Resampling

The use of stratified sampling procedures is another important way of removing biases in large-scale comparative studies. The idea here can most easily be described in terms of a demographic comparison of a particular characteristic between populations, for example, height in northern *vs* southern populations. It is possible that the overall population could be fractionated into further subdivisions on another parameter, potentially linked to height, say age (old *vs* young). In the above salt-bridge example, LOD statistics are analogous to computing the average height over the entire population regardless of age. However, the possibility that one population has more of a certain age group than another could potentially skew these statistics (eg Northerners are older and taller). To compensate for such bias in the sample one could take a representative sample from every age group and calculate the average height for that stratum. In the above case study, sets of 52 orthologous proteins present in each of the genomes were taken as a representative stratum. The strategy is illustrated in Figure 5c. Comparing results from this set with the genome-wide results supported the overall conclusion of salt-bridge prevalence in thermophile genomes.

Rank Statistics

Finally, rank statistics can be used to test the results of a comparative study. Rank ordering provides a more robust perspective of what is most abundant and what is rare. Therefore, if the rank of a certain event is consistently high, predominance of that event can be considered to be globally significant as opposed to just a 'local' effect arising out of a particular sequence bias. Furthermore, ranks provide a way of comparing disparate numerical values in a common framework.[88] In the above salt-bridge example, LOD values showed the prevalence of the 1–4 salt-bridge pair in comparison to other salt-bridge combinations in helices. It could be possible that the result was due to a certain group of proteins rich in salt bridges, and in the rest of the genome there were not that many salt bridges. Therefore, to validate the conclusion of comparative study, it is necessary to study the ranks of LOD values for all the helical pairs and compare them. If a pair is at the top of the ordered list of LOD values, then one could infer that this pair is among the most over-represented in the helices of the proteins for that organism. In the case study, ranks of salt-bridges in thermophiles were generally higher.

Comparison Based on Grouping Sequences into Pathways, Systems, and Beyond

In addition to sequenced-based and functional analysis, several genomic studies have analyzed genomes in terms of systems, specifically metabolic pathways and phylogenetic analysis. Similar to folds, metabolic pathways group together protein sequences. Since pathways are ordered clusters of sequences, their analyses can also reveal information about the physiology of the organism. Just as with folds one can cluster genomes based on the presence, absence or rank of a fold; one can group genomes based on whether or not they share a particular metabolic system. Furthermore, investigators working on microbial genomes have, through these investigations, created comprehensive metabolic maps.[147] Metabolic pathways can also be compared in terms of the properties of the enzymes and elementary modes.[148,149] Using metabolic networks' ●<aq1>●distances in pathways, one can measure and compare genomes based on the sequence information of enzymes and substrates in the pathway.[150] Pathways have also been analyzed by graph comparison methods where a pathway is considered as a graph with gene products as its nodes. This procedure leads to a formation of correlated clusters among the functionally related enzymes.[151] Any good analyses of metabolic networks based on genomic information requires substantial information with regard to networks, reactions and substrates.

There are several metabolic databases currently available. The KEGG database of metabolic pathways and regulatory pathways has a collection of approximately 100 metabolic pathways.[152] EcoCyc, specific to *E. coli*, has detailed information about the known metabolic pathways in *E. coli*. Studies of metabolic pathways can potentially help design new drugs for diseases caused by microbes and also help to understand how present drugs work within those pathways.

Beside metabolic pathways, there are other major areas of study where genomes are compared in terms of systems such as phylogenetic comparison, expression analyses in relation to various cellular functions, localization and events. Several new terms have been coined to describe them, such as proteomics, transcriptomics, metabolomics and pharmacogenomics. All these analyses give us a greater global knowledge with regard to the capabilities of systems such as metabolic pathways or transcription processes and their interrelationships.

CONCLUSION

There are many disparate methods that researchers use to compare genomes, from simple sequence comparison to protein structural comparisons to mRNA expression values. Each of these methods provides unique information with regard to genomes and how they compare or contrast. However, genome comparison based on protein structure is particularly advantageous as structures are well conserved between organisms even if the underlying sequence shows minimal homology. Also the relationship between structure and function is well defined. An important element of structural comparison between genomes is protein fold libraries that arrange the proteins into fold families. We discussed how different methods are used to build such libraries and how the concept of a parts list can be used to survey and re-survey the finite list of folds from an expanding number of perspectives. Genome-wide surveys are not limited to empirically defined structure, as structure predictions have proved to be fairly accurate in their predictive abilities. Moreover we discuss methods for, and underline the importance of, controlling for biases within a genome-wide study.

DUALITY OF INTEREST

None declared.

REFERENCES

1 Nowak R. Bacterial genome sequence bagged. *Science* 1995; **269**: 468–470.
2 Langreth R. Scientists unlock sequence of ulcer bacterium's genes. *Wall Street Journal* 1997 (Aug 7); B1.
3 Wade N. Thinking small paying off big in gene quest. *New York Times* 1997 02/03/97; Sect. A1.
4 Sherlock G. Analysis of large-scale gene expression data. *Curr Opin Immunol* 2000; **12**: 201–205.
5 Blaisdell BE, Campbell AM, Karlin S. Similarities and dissimilarities of phage genomes. *Proc Nat Acad Sci USA* 1996; **93**: 5854–5859.
6 Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 1995; **11**: 283–290.
7 Karlin S, Burge C, Campbell AM. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl Acids Res* 1992; **20**: 1363–1370.
8 Karlin S, Mrazek J, Campbell AM. Frequent oligonucleotides and peptides of the haemophilus influenzae genome. *Nucl Acids Res* 1996; **24**: 4263–4272.
9 Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997; **278**: 631–637.
10 Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol* 2000; **1**: ●●●●<aq2>●.
11 Koonin EV, Mushegian AR, Rudd KE. Sequencing and analysis of bacterial genomes. *Curr Biol* 1996; **6**: 404–416.
12 Brenner SE, Hubbard T, Murzin A, Chothia C. Gene duplications in *H. influenzae. Nature* 1995; **378**: 140.
13 Riley M. Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl Acids Res* 1997; **25**: 51–52.
14 Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997; **387**: 708–713.
15 Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* 1997; **274**: 562–576.
16 Tamames J, Casari G, Ouzounis C, Valencia A. Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* 1997; **44**: 66–73.
17 Teichmann SA, Park J, Chothia C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci USA* 1998; **95**: 14658–14663.
18 Nobusato A, Uchiyama I, Ohashi S, Kobayashi I. Insertion with long target duplication: a mechanism for gene mobility suggested from comparison of two related bacterial genomes. *Gene* 2000; **259**: 99–108.
19 Riley M. Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res* 1998; **26**: 54.
20 Green P. Ancient conserved regions in gene sequences. *Curr Opin Struct Biol* 1994; **4**: 404–412.
21 Koonin EV, Tatusov RL, Rudd KE. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc Natl Acad Sci USA* 1995; **92**: 11921–11925.
22 Ouzounis C, Kyrpides N, Sander C. Novel protein families in Archaean genomes. *Nucl Acids Res* 1995; **23**: 565–570.
23 Clayton RA, White O, Ketchum KA, Venter JC. The first genome from the third domain of life. *Nature* 1997; **387**: 459–462.
24 Debeljak N, Horvat S, Vouk K, Lee M, Rozman D. Characterization of the mouse lanosterol 14alpha-demethylase (CYP51), a new member of the evolutionarily most conserved cytochrome P450 family. *Arch Biochem Biophys* 2000; **379**: 37–45.
25 Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E. What's in a genome? *Nature* 1992; **358**: 287.
26 Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome iii. *Protein Sci* 1992; **1**: 1677–1690.
27 Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C *et al*. GeneQuiz: a workbench for sequence analysis. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA. AAAI Press: 1994, pp 348–353.
28 Casari G, Andrade M, Bork P, Boyle J, Daruvar A, Ouzounis C *et al*. Challenging times for bioinformatics. *Nature* 1995; **376**: 647–648.
29 Ouzounis C, Bork P, Casari G, Sander C. New protein functions in yeast chromosome VIII. *Protein Sci* 1995; **4**: 2424–2428.
30 Gaasterland T, Sensen CW. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* 1996; **78**: 302–310.
31 McClelland M, Wilson RK. Comparison of sample sequences of the *Salmonella typhi* genome to the sequence of the complete *Escherichia coli* K-12 genome. *Infect Immun* 1998; **66**: 4305–4312.
32 Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C *et al*. Automated genome sequence analysis and annotation. *Bioinformatics* 1999; **15**: 391–412.
33 Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A *et al*. Genome sequences and great expectations. *GenomeBiology.com* 2000; **2**: ●●●●<aq3>●.
34 Thornton JM, Orengo CA, Todd AE, Pearl FM. Protein folds, functions and evolution. *J Mol Biol* 1999; **293**: 333–342.
35 Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 1999; **288**: 147–164.
36 Gerstein M, Altman R. A structurally invariant core for the globins. *CABIOS* 1995; **11**: 633–644.
37 Gerstein M, Altman RB. Average core structures and variability measures for protein families: application to the immunoglobulins. *J Mol Biol* 1995; **251**: 161–175.
38 Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. *Nucl Acids Res* 1991; **19**: 6565–6572.

39 Henikoff S, Henikoff JG. Protein family classification based on searching a database of blocks. *Genomics* 1994; **19**: 97–107.
40 Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. Gene families: the taxonomy of protein paralogs and chimeras. *Science* 1997; **278**: 609–614.
41 Henikoff S, Pietrokovski S, Henikoff JG. Superior performance in protein homology detection with the Blocks Database servers. *Nucl Acids Res* 1998; **26**: 309–312.
42 Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN. The PRINTS protein fingerprint database in its fifth year. *Nucl Acids Res* 1998; **26**: 304–308.
43 Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 1995; **4**: 1618–1632.
44 Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1997. *Nucl Acids Res* 1997; **25**: 217–221.
45 Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA* 1994; **91**: 12091–12095.
46 Sonnhammer E, Eddy S, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997; **28**: 405–420.
47 Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl Acids Res* 1998; **26**: 320–322.
48 Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. *Nucl Acids Res* 1998; **26**: 323–326.
49 Fabian P, Murvai J, Hatsagi Z, Vlahovicek K, Hegyi H, Pongor S. The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nucl Acids Res* 1997; **25**: 240–243.
50 Sonnhammer ELL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci* 1994; **3**: 482–492.
51 Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. *Proc Natl Acad Sci* 1993; **19**: 6565–6572.
52 Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *Embo J* 1986; **5**: 823–826.
53 Chothia C, Gerstein M. Protein evolution. How far can sequences diverge? *Nature* 1997; **385**: 579, 581.
54 Jain KK. Genomics for business. *Drug Discov Today* 2001; **6**: 131–132.
55 Edwards A, Arrowsmith C, des Pallieres B. Proteomics: new tools for a new era. *Modern Drug Discovery* 2000; **5**: 35–44.
56 Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR *et al*. Structural proteomics of an archaeon. *Nat Struct Biol* 2000; **7**: 903–909.
57 Eisenstein E, Gilliland GL, Herzberg O, Moult J, Orban J, Poljak RJ *et al*. Biological function made crystal clear—annotation of hypothetical proteins via structural genomics. *Curr Opin Biotechnol* 2000; **11**: 25–30.
58 Murzin A, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins for the investigation of sequences and structures. *J Mol Biol* 1995; **247**: 536–540.
59 Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993; **233**: 123–128.
60 Johnson MS, Sali A, Blundell TL. Phylogenetic relationships from three-dimensional protein structures. *Meth Enz* 1990; **183**: 670–691.
61 Orengo CA, Flores TP, Taylor WR, Thornton JM. Identifying and classifying protein fold families. *Prot Eng* 1993; **6**: 485–500.
62 Pearl FM, Martin N, Bray JE, Buchan DW, Harrison AP, Lee D *et al*. A rapid classification protocol for the CATH domain database to support structural genomics. *Nucl Acids Res* 2001; **29**: 223–227.
63 Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. *Nucl Acids Res* 2000; **28**: 263–266.
64 Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucl Acids Res* 2000; **28**: 257–259.
65 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 1997; **25**: 3389–3402.
66 Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl Acids Res* 1998; **26**: 38–42.
67 Benson DA, Boguski M, Lipman DJ, Ostell J. *Genbank Nuc Acid Res* 1996; **24**: 1–5.
68 Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science* 1985; **227**: 1435–1441.
69 Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. *Proc Natl Acad Sci USA* 1988; **85**: 2444–2448.
70 Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics* 1998; **14**: 707–714.
71 Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* 1997; **273**: 349–354.
72 Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology: applications to protein modelling. *J Mol Biol* 1994; **235**: 1501–1531.
73 Baldi P, Chauvin Y, Hunkapiller T. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci* 1994; **91**: 1059–1063.
74 Eddy SR, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. *J Comp Bio* 1994; **9**: 9–23.
75 Taubes G. Software matchmakers help make sense of sequences. *Science* 1996; **273**: 588–590.
76 Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991; **253**: 164–170.
77 Eddy SR. Hidden Markov models. *Curr Opin Struc Biol* 1996; **6**: 361–365.
78 Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986; **188**: 415–431.
79 Staden R. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 1989; **5**: 89–96.
80 Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987; **84**: 4355–4358.
81 Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. *J Mol Biol* 1993; **232**: 1117–1129.
82 Bucher P, Karplus K, Moeri N, Hofmann K. A flexible motif search technique based on generalized profiles. *Comput Chem* 1996; **20**: 3–23.
83 Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. *Curr Opin Chem Biol* 2001; **5**: 51–56.
84 Sali A. Modeling mutations and homologous proteins. *Curr Opin Biotechnol* 1995; **6**: 437–451.
85 Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987; **326**: 347–352.
86 Bajorath J, Stenkamp R, Aruffo A. Knowledge-based model building of proteins: concepts and examples. *Protein Sci* 1993; **2**: 1798–1810.
87 Sali A, Sánchez R. Advances in comparative protein-structure modeling. *Curr Opin Struct Biol* 1997; **7**: 206–214.
88 Gerstein M, Hegyi H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev* 1998; **22**: 277–304.
89 Skolnick J, Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol* 2000; **18**: 34–39.
90 Chothia C. Proteins. One thousand families for the molecular biologist. *Nature* 1992; **357**: 543–544.
91 Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature* 1994; **372**: 631–634.
92 Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 1980; **136**: 225–270.
93 Gerstein M. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* 1998; **33**: 518–534.
94 Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* 2000; **10**: 808–818.
95 Ouzounis C, Kyrpides N. The emergence of major cellular processes in evolution. *FEBS Lett* 1996; **390**: 119–123.
96 Gerstein M, Lin J, Hegyi H. Protein folds in the worm genome. *Pac Symp Biocomput* 2000; ●●●<aq4>●: 30–41.
97 Sauder JM, Dunbrack RL Jr. Genomic fold assignment and rational modeling of proteins of biological interest. *Proc Int Conf Intell Syst Mol Biol* 2000; **8**: 296–306.

98 Fischer D, Eisenberg D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium. Proc Natl Acad Sci USA* 1997; **94**: 11929–11934.
99 Rychlewski L, Zhang B, Godzik A. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold Des* 1998; **3**: 229–238.
100 Mallick P, Goodwill KE, Fitz-Gibbon S, Miller JH, Eisenberg D. Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: validating automated fold assignment methods by using binary hypothesis testing. *Proc Natl Acad Sci USA* 2000; **97**: 2450–2455.
101 Dubchak I, Muchnik I, Kim SH. Assignment of folds for proteins of unknown function in three microbial genomes. *Microb Comp Genomics* 1998; **3**: 171–175.
102 Frishman D, Mewes H-W. PEDANTic genome analysis. *Trends Genet* 1997; **13**: 415–416.
103 Harrison PM, Echols N, Gerstein MB. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucl Acids Res* 2001; **29**: 818–830.
104 Honig B. Protein folding: from the levinthal paradox to structure prediction. *J Mol Biol* 1999; **293**: 283–293.
105 Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol* 1999; **9**: 368–373.
106 Finkel'shtein AV, Rykunov DS, Lobanov MI, Badretdinov FI, Reva BA, Skolnick J *et al*. [When and how can homologs overcome errors in the energy estimates and make the 3D structure prediction possible]. *Biofizika* 1999; **44**: 980–991.
107 O'Donoghue SI, Nilges M. Tertiary structure prediction using mean-force potentials and internal energy functions: successful prediction for coiled-coil geometries. *Fold Des* 1997; **2**: S47–52.
108 Hansson M, Gough SP, Brody SS. Structure prediction and fold recognition for the ferrochelatase family of proteins. *Proteins* 1997; **27**: 517–522.
109 Rost B. PHD: predicting one-dimensional protein secondary structure by profile-based neural networks. *Meth Enz* 1996; **266**: 525–539.
110 Defay T, Cohen FE. Evaluation of current techniques for *ab initio* protein structure prediction. *Proteins* 1995; **23**: 431–445.
111 Pedersen JT, Moult J. *Ab initio* protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins. *Proteins* 1997; Suppl **1**: 179–184.
112 Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Meth Enzymol* 1996; **266**: 540–553.
113 Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978; **120**: 97–120.
114 Gibrat JF, Garnier J, Robson B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J Mol Biol* 1987; **198**: 425–443.
115 King RD, Saqi M, Sayle R, Sternberg MJ. DSC: public domain protein secondary structure predication. *Comput Appl Biosci* 1997; **13**: 473–474.
116 Livingstone CD, Barton GJ. Identification of functional residues and secondary structure from protein multiple sequence alignment. *Meth Enzymol* 1996; **266**: 497–512.
117 Gerstein M, Levitt M. A structural census of the current population of protein sequences. *Proc Natl Acad Sci U S A* 1997; **94**: 11911–11916.
118 Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* 1986; **15**: 321–353.
119 Gribskov M, Devereux J. *Sequence Analysis Primer*. Oxford University Press: New York, 1992.
120 Kyte J, Doolittle RF. A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 1982; **157**: 105–132.
121 Jähnig F. Structure predictions of membrane proteins are not that bad. *TIBS* 1990; **15**: 93–95.
122 von Heijne G. Membrane proteins: from sequence to structure. *Annu Rev Biophys Biomol Struct* 1994; **23**: 167–192.
123 von Heijne G. Principles of membrane protein assembly and structure. *Prog Biophys Mol Biol* 1996; **66**: 113–139.
124 Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Functional & Integrative Genomics* 2000; **1**: 76–88.
125 Auerbach G, Ostendorp R, Prade L, Korndorfer I, Dams T, Huber R *et al*. Lactate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: the crystal structure at 2.1 A resolution reveals strategies for intrinsic protein stabilization. *Structure* 1998; **6**: 769–781.
126 Hennig M, Darimont B, Sterner R, Kirschner K, Jansonius JN. 2.0 A structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: possible determinants of protein stability. *Structure* 1995; **3**: 1295–1306.
127 Knapp S, de Vos WM, Rice D, Ladenstein R. Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium *Thermotoga maritima* at 3.0 A resolution. *J Mol Biol* 1997; **267**: 916–932.
128 Hennig M, Sterner R, Kirschner K, Jansonius JN. Crystal structure at 2.0 A resolution of phosphoribosyl anthranilate isomerase from the hyperthermophile *Thermotoga maritima*: possible determinants of protein stability. *Biochemistry* 1997; **36**: 6009–6016.
129 Korndorfer I, Steipe B, Huber R, Tomschy A, Jaenicke R. The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 A resolution. *J Mol Biol* 1995; **246**: 511–521.
130 Russell RJ, Ferguson JM, Hough DW, Danson MJ, Taylor GL. The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 A resolution. *Biochemistry* 1997; **36**: 9983–9994.
131 Salminen T, Teplyakov A, Kankare J, Cooperman BS, Lahti R, Goldman A. An unusual route to thermostability disclosed by the comparison of *Thermus thermophilus* and *Escherichia coli* inorganic pyrophosphatases. *Protein Sci* 1996; **5**: 1014–1025.
132 Spassov VZ, Karshikoff AD, Ladenstein R. The optimization of protein-solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. *Protein Sci* 1995; **4**: 1516–1527.
133 Szilagyi A, Zavodszky P. Structural basis for the extreme thermostability of D-glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima*: analysis based on homology modelling. *Protein Eng* 1995; **8**: 779–789.
134 Wallon G, Yamamoto K, Kirino H, Yamagishi A, Lovett ST, Petsko GA *et al*. Purification, catalytic properties and thermostability of 3-isopropylmalate dehydrogenase from *Escherichia coli. Biochim Biophys Acta* 1997; **1337**: 105–112.
135 Yip KS, Stillman TJ, Britton KL, Artymiuk PJ, Baker PJ, Sedelnikova SE *et al*. The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* 1995; **3**: 1147–1158.
136 Kawamura S, Tanaka I, Yamasaki N, Kimura M. Contribution of a salt bridge to the thermostability of DNA binding protein HU from *Bacillus stearothermophilus* determined by site-directed mutagenesis. *J Biochem (Tokyo)* 1997; **121**: 448–455.
137 Mande SS, Gupta N, Ghosh A, Mande SC. Homology model of a novel xylanase: molecular basis for high-thermostability and alkaline stability. *J Biomol Struct Dyn* 2000; **18**: 137–144.
138 Hartley BS, Hanlon N, Jackson RJ, Rangarajan M. Glucose isomerase: insights into protein engineering for increased thermostability. *Biochim Biophys Acta* 2000; **1543**: 294–335.
139 Qu CC, Akanuma SS, Tanaka NN, Moriyama HH, Oshima TT. Design, X-ray crystallography, molecular modelling and thermal stability studies of mutant enzymes at site 172 of 3-isopropylmalate dehydrogenase from *Thermus thermophilus. Acta Crystallogr D Biol Crystallogr* 2001; **57**: 225–232.
140 Xiao L, Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins. *J Mol Biol* 1999; **289**: 1435–1444.
141 Vetriani C, Maeder DL, Tolliday N, Yip KS, Stillman TJ, Britton KL *et al*. Protein thermostability above 100 degrees C: a key role for ionic interactions. *Proc Natl Acad Sci U S A* 1998; **95**: 12300–12305.
142 Lebbink JH, Knapp S, van der Oost J, Rice D, Ladenstein R, de Vos WM. Engineering activity and stability of *Thermotoga maritima* glutamate dehydrogenase. I. Introduction of a six-residue ion-pair network in the hinge region. *J Mol Biol* 1998; **280**: 287–296.
143 Scholtz JM, Qian H, Robbins VH, Baldwin RL. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry* 1993; **32**: 9668–9676.

144 Huyghues-Despointes BM, Scholtz JM, Baldwin RL. Effect of a single aspartate on helix stability at different positions in a neutral alanine-based peptide. *Protein Sci* 1993; **2**: 1604–1611.
145 Russell RB, Barton GB. Multiple protein sequence alignment from tertiary structure comparisons. Assignment of global and residue level confidences. *Proteins* 1992; **14**: 309–323.
146 Grindley HM, Artymiuk PJ, Rice DW, Willett P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol* 1993; **229**: 707–721.
147 Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res* 1998; **8**: 203–210.
148 Galperin MY, Koonin EV. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica* 1999; **106**: 159–170.
149 Dandekar T, Schuster S, Snel B, Huynen M, Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J* 1999; **343**: 115–124.
150 Forst CV, Schulten K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J Comput Biol* 1999; **6**: 343–360.
151 Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucl Acids Res* 2000; **28**: 4021–4028.
152 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucl Acids Res* 2000; **28**: 27–30.