Comparison of sequences, protein 3D structures and genomes

László KAJÁN[1], Kristian VLAHOVICEK[1,2], Oliviero CARUGO[1,3], Vilmos ÁGOSTON[4], Zoltán HEGEDÜS[4] and Sándor PONGOR[1]

[1]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy*

[2]*Molecular Biology Department, Biology Division, Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia*

[3]*Department of General Chemistry, Pavia University, viale Taramelli 12, 27100 Pavia, Italy*

[4]*Bioinformatics Group, Biological Research Center, Hungarian Academy of Sciences, Temesvári krt. 626726 Szeged, Hungary*

Abstract. The analysis of similarity is a fundamental task in comparing sequences, three dimensional structures as well as genomes and molecular networks. This chapter reviews the common principles underlying these diverse applications.

Introduction

The basic concepts of similarity analysis – as presented in the first part of this review – provide a common framework for the classification of newly identified the protein sequence or protein 3D structure. Classification of an object implies placing it into the already existing categories or marking it as "unknown" i.e. as a potential initiator of a new category. This process usually consists of the following steps.

Recognition of similarity. This is a qualitative decision that is often based on some approximate quantitative measure. In sequence analysis, if the raw alignment score is above a threshold, the similarity is considered significant and retained for further analysis. In the case of protein 3-D structures the preliminary evaluation is often based on visual inspection.

Next, the basis of similarity, i.e. a common substructure is identified. This is carried out by matching of the equivalent entities and relationships, and sequence alignments as well as structural alignments are the best examples. Determination of matching by computers involves maximization of a similarity measure (or minimization of a distance measure), and the final value of the respective parameters is used as a numeric measure of similarity.

Evaluation of similarity. First a decision has to be made whether or not the similarity is biologically important, and the protein is either assigned to a known similarity group or it will be considered as the initiator of a new group. This decision is usually based on one or more similarity scores as well as on the alignment, but human judgment is hard to replace and at this stage.

Representation of similarity in databases. Once the similarity is established, it has to be added to the annotation of the protein in the sequence and or 3-D databases. Protein superfamilies, structural domains, orthologous groups etc. are determined by similarity analysis, and there is large number of secondary databases that are dedicated to the curation

of the underlying similarity groups. Apart from narrative descriptions there are two general avenues to describe similarity groups. Cladograms are classifications that can be established using proximity measures and represent the internal structure of the similarity group. Common patterns on the other hand are usually derived from alignments and represent common substructures present in the members of the similarity group.

The above steps are not always obvious for the users. For example, sequence similarity search programs present the results corresponding to step II, while some of the 3-D similarity search servers provide only a qualitative suggestion corresponding to step I. What is apparent however that all methods include a preliminary, approximate estimation of similarity, followed by a filtering and finally an alignment step.

This section provides a brief overview of how similarity scoring in used in the comparison of sequences, protein 3-D structures and entire genomes. In these fields, similarity measures are used for database searching, for classification and for phylogenetic analysis. A comprehensive overview of these broad fields would be far beyond the scope of this chapter. Instead, we will attempt to highlight, using the terminology introduced in the previous sections, the common themes underlying these three diverse areas.

1. Sequence comparison

Sequences are the simplest descriptions of macromolecules that use residues (amino acids, nucleotides) as entities and sequential vicinity as the only relationship between them. Sequence comparison algorithms use essentially the same principle for similarity scoring. The simple proximity measure is related to the Hamming distance (i.e. no gaps allowed, as shown in Fig. 3.2). The scoring matrices used in DNA as well as protein comparisons are constructed in such a way that similar residues give high scores, so the resulting measure can be called a Hamming similarity measure, rather than a distance. The optimizable substructure similarity is the string similarity measure (equation 5) in which the position and number of the gaps as well as the range of alignment is determined by optimization. The result is a maximal matching, and the alignment score is a *local* or *global* maximum value depending on the algorithm used. Algorithms of global alignment (Needleman-Wunsch, [1]) or local alignment (FASTA [2], BLAST [3],) have been the subject of several excellent, recent reviews [see, e.g. [4,5]], the current section focuses on the principles of scoring, i.e. how a similarity score is transformed into a probabilistic measure.

We will use a simple classification: *General methods* of comparison use a general statistical description of random similarities for calculating the significance value to alignment scores. *Specific methods* use application-specific descriptions of the biologically important target groups, such as protein families, domain sequence groups etc. These groups are often too small for statistics, so specific methods rely instead on additional, *a priori* knowledge.

The most frequently used general methods (BLAST [3], FASTA [2], Smith-Waterman [6]) are based on local sequence alignment. The resulting sequence similarity scores do not preserve the metric properties (can not be converted into metric distances), on the other hand they have the advantage that the distribution of random similarities can be described in an analytical form. This is because scores are maximal values, and the maximum of a large number of independent identically distributed (i.i.d) random variables tends to an extreme value (or Gumbell) distribution, just as the sum of a large number of i.i.d. random variables tends to a normal distribution [7]. The underlying statistics was described in detail by Karlin and Altschul [8,9] for the *BLAST* program. Originally, BLAST used local alignments without gaps called high-scoring segment pair (or *HSP*), in which scores were maximized in the sense that they could not be further improved by extension or

trimming. We will use HSPs as an example, adding that the description of gapped BLAST, FASTA and Smith/Waterman scores follows a similar statistics. The random emergence of *HSP*s was studied on random sequences in which the occurrence amino acid residues is independent, with specific background probabilities for the various residues. For two sufficiently long (*m* and *n*) sequences, the expected number of HSPs with score at least *S* is given by the formula

[1] $$E = Kmne^{-\lambda S}$$

where K and λ are constants that can be considered a can be as natural scales for the search space of size $m \cdot n$ and the scoring system. The raw score *S* is defined by a formula given in figure x. The number of random HSPs with score $\geq S$ is described by a Poisson distribution and the probability of finding at least one such HSP is

[2] $$P = 1 - e^{-E}$$

P is the statistical significance, the probability of finding a score *S* (or bigger) by chance. It is important to note that this simple statistics is also approximately valid for gapped alignments used by modern alignment programs, and this makes it possible to give a more objective, probabilistic interpretation to similarity scores.

Global alignments are found via an exhaustive search for the maximal matching between two sequences, based on such methods as the Needleman-Wunsch algorithm [1]. Global alignment scores can be transformed to metric distance scores, which is important for clustering. On the other hand, very little is known about the random distribution of optimal global alignment scores, so a rigorous probabilistic interpretation is not possible in this case. A practical approach is based on generating many random sequence pairs of the appropriate length and composition, and calculating the optimal alignment score for each. The average S_r and the standard deviation σ_r of the random scores can then be compared with original score *S* score, and a Z score

[3] $$Z = \frac{|S - S_r|}{\sigma_r}$$

can be used as an approximate measure of significance. Namely, even though Z resembles the Student *t* value, but rigorously speaking it cannot be converted into a *P* value since the underlying distribution is not a normal distribution. Only an approximate interpretation is thus possible, for example if 100 random alignments have scores inferior to the alignment of interest, the *P*-value in question is likely less than 0.01. It is important to note that the meaning of this statistics is different from the one derived from a database of random similarities (equation 16). Namely, for two sequences of similar, but unusual amino acid composition, the Z-score may be a low value, even is the two sequences compared are both very different from the rest of the database.

The general methods of sequence comparison can be used to divide the sequence database into clusters. In principle, a metric distance measure (such as can be derived from global alignment scores) is a prerequisite for statistical clustering. Given the large size of databases, both global alignments and statistical clustering methods are compute-intensive. On the other hand, the protein sequence space is sparsely populated and the existing natural sequences form well-separated clusters, which makes it possible to use efficient, approximate methods for clustering. Krause and Vingron used a threshold-based, iterative procedure based on BLAST for identifying consistent protein clusters [10,11]. The result an objective picture of the sequence space in terms of similarities, but the clusters have to be

compared with knowledge based groups, such as protein families etc. With this approach, protein domains that are shared among several protein families lead to the merging of protein family clusters.

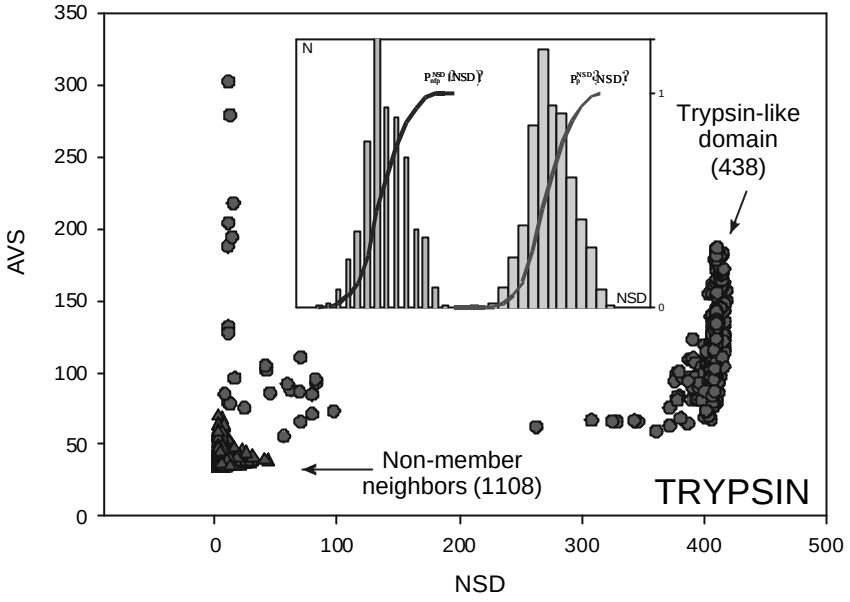
A sharp distinction between biologically significant and random similarities is not possible from the scores alone – such decisions still require *a priori* knowledge, namely biological knowledge (e.g. knowledge of the overall domain structure of the protein, the exon-structure of the genes) as well as a knowledge of the previously known similar sequences. In addition to the general methods of sequence comparison mentioned above, there are a number of dedicated specific methods, based on some explicit representation of biologically important similarity groups such as protein domain sequences. A sequence similarity group can be represented by a consensus description that represents e.g. a sequence pattern that is shared by all members of the group. As such patterns can be obtained by multiple sequence alignments, there is a large variety of algorithms that represent multiple alignments in terms of consensus sequences, regular expressions, position-specific scoring matrices or profiles, hidden Markov models (HMMs) or neural networks (for recent reviews see [5,12]). These consensus descriptions can then be used to decide whether or not a new query sequence is member of a given similarity group. The similarity measures used to compare a query with these representations are similar to the ones described in this review, the details can be found in the original publications as well as the reviews cited above.

Another group of specific approaches uses a graph-theoretical representation of similarity groups, which is an exemplar-based description. Sequences within a similarity group are related to each other by specific similarity (Figure 3.1.), for example each member of the group is related to at least one other member with a similarity score greater that a certain threshold [13]. Protein domains are typical examples of well-defined similarity groups. On the other hand, many of the known proteins are composed of modules, so the score determined between two such proteins will express the similarity of the building blocks, rather than that of the two proteins.

The similarities of protein domain groups can be defined on a threshold basis. In the SBASE protein domain sequence library, a sequence is considered as member of a domain group if it is similar to at least NSD_t members of the group, with an average similarity score of AVS_t where NSD_t and AVS_t are threshold values automatically determined from a database vs. database comparison with the BLAST program. A later extension of this scoring system takes into consideration the distribution of similarity scores in the neighborhood of each similarity group and uses a probabilistic score. For each raw scores, four probability values are read from the precomputed distributions shown in **Figure 1**, and the score is derived from the sum of these distributions [14]. From the computational point of view, this approach is similar to the memory-based computing paradigm [15], the memory of the system is a database vs. database comparison [16,17].

The approach underlying the COG (Clusters of Orthologous Sequences) databank is based on grouping sequences together that are mutually the nearest neighbours of each other in terms of sequence similarity score [18]. Such tight groups or cliques can be extended to larger similarity groups, which is the basis of identifying orthologous proteins. This approach is especially successful in prokaryotic genomes in which multidomain proteins are not abundant.

Recent approaches combine many of the previous concepts. The underlying philosophy is that database search results should contain all information necessary to find distant similarities – such as the weak similarities of protein domains – and that these might be found via a clever sorting of the search results. Namely, the alignment scores (an the *P* values) traditionally used to sort the result constitute only one dimension of the sorting.



$$C = TP_{NSD} + P_p(NSD), + P_{nfp}(NSD), + TP_{AVS} + P_p(AVS) + P_{nfp}(AVS),$$

Figure 1. The principle of classifying domains in SBASE [14] (See text for explanations).

Alignments can be sorted according to their position within the query, as well as according to their common sequence patterns. Recent versions of BLAST, incorporate position specific scoring known from profile methods (PSI-BLAST) as well as pattern-specific searches (PHI-BLAST) [19-21].

Given the ease and speed of current sequence alignment algorithms, approximate methods based on unstructured descriptions are used only in specific applications. In composition-based methods, the sequences are described as vectors, in terms of the amino acid, dipeptide, tripeptide etc. composition, and the comparison is based on simple distances such as the Euclidean distance. Same as with other unstructured descriptors, the calculation is very fast, especially since the database can be stored in the form of pre-calculated vectors. The number of vector components (the resolution of the description) has to be selected with care, and this is done either heuristically, or using an algorithm to automatically select and/or weight those amino acid words that give the best separation between a test group and a control group. In this manner group-specific distance functions can be developed. The resolution of the description can be fine-tuned e.g. by decreasing the amino acid alphabet (to 4,5, etc. letter alphabets instead of 20) and or by increasing the word size (dipeptides, gapped dipeptides, tripeptides etc.). Examples include the composition-based protein sequence search of Hobohm and Sander [22], as well as the promoter-search program of Werner et al. [23-25]. Simple applications include the recognition of coding regions based on codon-usage. Composition-based methods are very useful for building recognizers for any sequence group for which a sufficient number of examples are known. Given a test group and a control group of sequences, one can compare the frequency of arbitrary words (provided as a list) between these two groups. The most characteristic words can be selected based on simple measures such as the Mahalanobis distance, and used for recognizing potential new members of the test group [26]. Similar algorithms are often used in gene prediction systems [27].

Distributions are less frequently used for representing sequences, even though methods of comparing sequence profiles such as hydrophobicity plots, secondary structure propensity plots were developed already in the 1980-es. Fourier transforms of

hydrophobicity plots have been used to recognize amphipathic helices as well as to build classifiers to various protein groups. A review on these applications is in [28].

2. Comparison of 3D structures

Comparison of 3D structure is used in a variety of fields such as fold recognition, structural evolution studies and drug design, and the protocols are as diverse as the fields themselves. E.g. in the comparison of 3D structures produced on the same protein molecule by NMR methods, all the equivalent atom-pairs are a priori known and can be used in the comparison. In contrast, determination of folds is based on the backbone C? atoms only and the equivalences have to be determined by the calculation itself. In this section we will briefly summarize the similarity/distance functions used for backbone comparison, concentrating on the similarity/distance measures used rather than the goal and/or implementation of the actual algorithms. In the majority of the cases, the approach used for structural alignments is quite similar to that used in sequence analysis (finding alignment paths in a distance matrix or optimizing the range by successive omission or additions). This is because 3D structures can be compared in terms of their (overlapping) peptide fragments, and a series of peptide fragments is a linear, sequence-like representation. For example, one can compute an *rmsd* between the peptide fragments of two proteins and construct a distance-matrix with the resulting values [29,30]. But there are many ways to represent peptide fragments as vectors, and then one can use any of the vector-distance formulas to produce the values of the distance matrix. For example, vectors of torsional angles [31,32], curvature and torsion parameters of peptide fragments [33,34] have been used by early comparison methods, as reviewed by Orengo [35]. More recent methods include structural alphabets described in terms of dihedral angles [36,37] or on distance geometry [38,39]. In the latter method, the size of the alphabet (the minimum number of fragments necessary to describe the observed data) is 27 derived from statistical optimisation. The similarity search is then carried out by Smith-Waterman alignment.

The similarity measures described in this section can be classified according to the use of atomic (residue-based) descriptions, or higher-order descriptions such as secondary structure elements. Another important difference is that some of the methods can be used to produce structural alignments while others are only preliminary filters indicating similarity without providing a structural alignment.

Methods based on superposition of atoms use the *rmsd* distance (section x, above) Even though the results of atom superposition methods are generally considered superior to most computational alternatives, and very low *rmsd* values are indicative of identical structures – rmsd can be used only with caution as a quantitative indicator of similarity. In addition, there is no accepted and reliable statistical model that would allow to use *rmsd* as a probabilistic score with a statistical significance, moreover *rmsd* does not penalize gaps. Therefore there a number of alternative similarity scores have been developed for obtaining optimal structural alignments even though the final results are always characterized in terms of the rmsd score.

One group of similarity scores is based on vectors or sets of vectors assigned to each position within a protein structure. The parameters of the vector represent various features. Methods developed by Taylor and Orengo [40,41] assigned a set of intramolecular $C_?$?$C_?$ vectors to each residue position, or used various geometric features as parameters of the vector assigned to each residue position. As a result, a protein structure was converted into a series of residue vectors, and two structures could be compared to give a so-called residue matrix in which the elements are calculated as a vectorial difference (city-block

distance of vectors, equation [2]). The optimal structural alignment can be determined by a dynamic programming algorithm.

A roughly similar approach was used by Holm and Sander for the very popular DALI server [42]. In the underlying method the $C_?$ atoms are characterized by vectors the parameters of which are the elements of distance matrix. The local vectors are then compared in terms of residue similarity scores such as

[4] $$?^{R}(i, j) ? ?^{R} ? \left| d_{ij}^{A} ? d_{ij}^{B} \right|$$

or

[5] $$?^{E}(i, j) ? \begin{matrix} ? \\ ? \\ ? \end{matrix} ?^{E} ? \frac{\left| d_{ij}^{A} ? d_{ij}^{B} \right|}{d_{ij}^{*}} \begin{matrix} ? \\ ? \\ ? \end{matrix} e^{?(d_{ij}^{*})^{2}/?}$$

The subscript A,B refer to residues in structure d_{ij} are the elements of the hexapeptide distance matrices i.e. elements of the residue vectors. d_{ij}^{*} denotes the average of d_{ij}^{A} and d_{ij}^{B}, $?^{R}, ?^{E}$ and ? are constant. A and B, respectively. Superscript *R* denotes rigid comparison [eqn. 4], E refers to an elastic comparison dampened by a negative exponential term [eqn.5]. As can be seen, summing the residues similarity measures $?^{R}$ or $?^{E}$ results in quantities related to the city block distance. Comparison of two proteins A and B is then carried out using a distance matrix whose elements are equal to either $?^{R}(i, j)$ or $?^{E}(i, j)$, where i and j refer to two pairs of structurally aligned residues: i(A), i(B), j(A), and j(B). The optimization task is to find the best set of equivalences between A and B that maximize this function and the structural alignment is obtained by an optimization algorithm (Monte Carlo optimization) To improve convergence, various heuristics are used to obtain a reasonable starting point.

The residue similarity score of Levitt and Gerstein [43] has the formula

[6] $$S_{i,j} ? M /(1? (d_{ij} / d_0)^2)$$

where d_{ij} is the distance between $C_?$ atoms of the two structures compared, *M* and d_0 are constants. S_{ij} values are elements of a similarity matrix from which an optimizeable substructure similarity measure S_{str} can be calculated by introducing gaps. The S_{str} score is defined as

[7] $$S_{str} ? M(? _{ij} 1/(1? (d_{ij} / d_0)^2) ? N_{gap} / 2)$$

The structural alignment is carried out with a dynamic programming method such as the Smith-Waterman algorithm. Levitt and Gerstein found that random structural similarities determined by this method follow the same extreme value distribution as BLAST scores (or Smith-Waterman sequence alignment scores), so the results can be characterized in terms of *P* values [43].

As superposition methods are compute intensive, a number of simplified representations have been developed. One general strategy is to represent the protein by a set of secondary structure elements (SSEs), characterized by their position within the polypeptide sequence and the position in 3D space and are usually represented as vectors fit to the $C_?$ atoms. This is another kind of entity-relationship description in which SSEs are the nodes and a variety of parameters (such as distances, angles ec) are used to describe relationships. The rationale is that superposition of a few SSEs is less compute intensive

than superposing a large number of $C_?$ atoms, so one can use algorithms that could not cope with large atomic detail structures. In addition, SSEs incorporate added knowledge on molecular geometry. The success of the process depends on i) how secondary structures are assigned; ii) how the similarity between two secondary structural elements of two proteins is estimated; iii) how the overall similarity between the two proteins is defined.

Although the SSEs (at least the most common like helices and strands) are clearly defined, different assignment result from different assignment algorithms [44-46]. Consequently, different representations of the protein structures may arise. A further problem is which SSE types are considered. Very often a two-states classification is used: helix, including 3/10 and pi, and stand. There are nevertheless exceptions. Orengo et al. [44-46], for example, adopt a three-states classification: alpha-helix, 3/10-helix, and strand.

The similarity between secondary structural elements in two proteins is usually estimated by comparing each *pair of SSEs* of one protein with each pair of the other. The 3D arrangement of a two secondary structural elements in a protein is usually defined by their distance, their plane angle, and their torsion. A similarity score can then be computed for each pair of two secondary structural elements. The resulting matrix of similarity scores can then be scrutinized with dynamic programming techniques [41,47-49], treated as a maximum clique problem [50], with pseudo-distance matrices [51], or with cluster analysis [52]. The alignment of the secondary structural elements is eventually followed by a superposition of the $C_?$ atoms with an initial structural alignment that depends on the secondary structure alignment. The overall similarity between the two structures can be then estimated on the basis of the *rmsd* values [50] of with more sophisticated figures of merit that considers also the quality of the secondary structure fit.

The fragment-pair approach is also amenable to probabilistic interpretation. The VAST program of Bryant and coworkers [53,54] provides BLAST-like *P* significance values. VAST's elementary unit of comparison is a simplified *rmsd* score resulting from a superposition of the endpoints of SSE pairs "trimmed" to the same length. First *rmsd* values are converted into log-odds scores using precomputed values of comparison of SSE pairs from related and unrelated structures, then a combined score S_o is calculated from the i best SSE pairs found to mattch between the query and a database entry. The principle of converting S_o into a *P* value is similar to that used by BLAST, given in equations. 15-17, but relies on tabulated statistics, rather then on analytical formulae. Let the probability of finding a substructure of size i with a score $S_i?S_o$ be denoted as $P(S_i?S_o)$. In VAST, the value of $P(S_i?S_o)$ is estimated as a function of i and S_i, using tabulated values resulting from random comparisons. The expected number E of finding at least one score $S_i?S_o$ by chance will also depend on the size of the search space which can be defined as the total number of possible common substructures of i SSEs between the two proteins, a number denoted by N_i. The equation computed by VAST is then

[8] $$E ? ?_i N_i P(S_i ? S_o)$$

The sum is calculated for all i values using the tabulated $P(S_i?S_o$ values. Same as with BLAST, if E is small (e.g. E<0.01) it is also a *P* value. The method is very fast, due to the precomputed statistics, and accessible at the NCBI web site.

A variety of other procedures that represent the protein 3-D structure as an ensemble of secondary structural elements have also been proposed. In Martin's approach [55], secondary structural elements are given one of the letters of an alphabet that identify the secondary structure type, direction, length, and solvent accessibility. Two proteins can be thus compared with the simple Needlemann-Wunsch algorithm. Murthy [56] used dynamic programming techniques to optimally superpose secondary structural elements.

Mitchel [57] developed a graph-like representation, using secondary structural elements as nodes, and angles and distances as edges, the largest common substructure was then identified by subgraph isomorphism algorithm developed by Grindley et al. [58]. Harrison and associates [59] further developed this approach and introduced a similarity measure S_{grath} based on the number of SSEs and residues in two proteins and in the largest common subgraph:

[9] $$S_{grath} = \frac{5}{8}\sqrt{\frac{CS}{SS1}\frac{CS}{SS2}} \cdot \frac{Min(R1,R2)}{Max(R1,R2)} \cdot 2\sqrt{\frac{CR1}{R1}\frac{CR2}{R2}}$$

where the two proteins compared have *SS1* and *SS2* secondary structures and *R1* and *R2* amino acid residues, respectively, their comparison generates a largest clique of size *CS*. The largest clique is produced from a set of secondary structures that contain a total of *CR1* and *CR2* residues in protein 1 and protein 2, respectively. This similarity measure is reported to be independent of fold size and was used to characterize the fold space represented by the CATH database [59].

Finally, there are methods that do not use superposition but define simple similarity scores instead. PRIDE [60] is based on the distribution of intramolecular C_α-C_α distances incorporated into a set of histograms for C_α pairs separated by 3 to 30 residues. The comparison between two proteins is thus reduced to the comparison of 28 distribution pairs, which can be carried out by a standard statistical method of contingency table analysis and yields a probability value. The average value of these 28 single similarity scores was defined as the Probability of identity or PRIDE score [60]. Pride has a value between 0 and 1, and has metric properties which makes it suitable for clustering large datasets. The calculation is extremely fast (perhaps the fastest available today), database search and fold assignment, clustering of structures are possible on line. When used as a simple nearest neighbor classifier, PRIDE reaches 99.5% success in fold recognition, based on the C,A,T, H classes of the CATH database. This method is available via a web server at ICGEB.

Another recent, fast comparison method by [61] uses a vector representation of protein folds which is based on topological invariants called Gauss integrals, each representing a topological property of the backbone space curve [62]. 30 such integrals are calculated for two proteins, which are then compared in terms of a 30 dimensional Euclidean distance. A classifier built on Gauss integrals has a reported accuracy of 96.8% on the C,A,T classes of the CATH database [61].

3. Genomes, proteomes, networks

Designing representations for genomes, proteomes and networks is a real challenge and we are only at the first steps of this new era. The representations in current use follow the entity-relationship tradition, for example genomes are represented as linear array of genes and other DNA segments. The entities – genes – are predicted with gene-prediction programs or are determined experimental methods, and this adds a new layer of knowledge to the molecular data. The relationships are manifold but are predominantly binary in nature. Examples of relations include physical vicinity, distance along the chromosome, regulatory links extracted from DNA chip data and so on. The resulting picture is a graph of several ten-thousand nodes and relatively few edges per node denoting various relationships. The description of proteomes is only somewhat different. The proteins are described in functional, biochemical and structural terms, and the relationships between proteins include metabolic relationships (sharing substrates in metabolic pathways) as well as structural relationships (sequence and structural similarities). Even this sketchy

introduction implies that we deal with new a kind of complexity that originates, from the numerous and to a large extent, unknown interactions between the molecules. On the other hand, the study of large networks – such as Internet, social- road- and electric networks, etc. – has provided interesting insights that have been successfully applied to genomes, proteomes and bibliographic networks.

From the computational point of view, genomes and proteomes are described as very large graphs in which the nodes (genes, proteins) and the edges (relations) are unknown or unsure. These large and fuzzy descriptions are in sharp contrast with the descriptions developed for well-defined molecular structures, but the methods are not dissimilar to those used in other fields. Given the large and different genome sizes as well as the uncertainties of the data, structured descriptions are not very useful for comparison. Simple, unstructured representations like sets or vectors that can be easily compared in terms of their known components are widely used. The approaches differ how the components are selected and compared. (It is noted that this section concentrates only on those genome-comparison studies that use genome-level descriptions. For phylogenetic approaches to genome comparison see [63,64].

One group of approaches use predefined components, given in the form of a classification. Proteins can be classified into several thousand orthologous groups (COGs) and a genome (proteome) can be described by as a vector with a corresponding number of components, each component denoting the presence or absence of a given protein group [63]. This is an extremely simplified unstructured description, but the selected components of the vector adequately describe the entire universe of protein functions as we know it today. Two such vectors can then be compared using the Jaccard coefficient, and a related distance measure (1 minus the coefficient) can be used as a metric for classifying the genomes. This is a fast procedure that has no adjustable parameters, nevertheless it gave results in good agreement with other, more subjective methods. In a similar way, proteins can be grouped according to their similarity to sequences representing known 3D folds. In a similar manner, the genomes can be classified in terms of the 3D folds [65,66].

Metabolic data are a further example of predefined component classification that can be used for genome comparison [67,68]. A proteome can be described in terms of the constituting enzymes, substrates, intermediate complexes such as given in the WIT database [69]. Organism data can then be converted into vectors representing enzymes or substrates in pathways or pathway-groups. For example, in a vector representing the metabolic pathways of *E. coli* in terms of enzymes, a parameter e_i is an integer denoting the number of times enzyme i occurs in the metabolic pathways of the organism. Such vectors can then be compared using any of the vector similarity/distance measures described above. A classification of genomes based on vectors representing the metabolic and information-processing pathways in terms of enzymes and substrates has shown that the system-level organization of *Archea* and *Eukarya* are similar [70]. This comparison was based partly on presence/absence data and the Jaccard coefficient, and partly on comparing the ranking data of component frequencies.

Another group of representations uses sequence comparison to dynamically define matching components between two genomes. The matching pairs of genes can be selected based on BLAST scores [71], Smith-Waterman scores [72,73]. The intergenomic distances can then be based on the list of shared (as well as total) components present in two genomes, using e.g. the Jaccard coefficient. A particularly interesting version of this method uses vicinal gene-pairs with conserved direction of transcription, identified from Smith-Waterman searches [74-76]. Given the matching vicinal pairs in the two genomes as substructures, one can proceed in the usual way. This method thus preserves the speed of the comparison but uses substructures that are richer in detail i.e. capture a part of the gene order.

Unfortunately, the gene-based substructures cannot be increased without practical limits: at present, quantitative genome comparisons are seemingly limited to gene (protein) pairs. On the other hand, higher-order patterns can be very informative in the qualitative sense. Studies on conserved local gene-order revealed that in addition to the known operons, there are larger units – über-operons or super-operons – that are conserved in terms of functional and regulatory context [77,78]. This takes us to a familiar world of known patterns: operons and metabolic pathways are both higher order patterns (directed graphs) defined in terms of entities and relationships. Comparison of related metabolic pathways is a subgraph isomorphism problem [79], and related techniques underlie the Kyoto encyclopedia of genes and genomes [80].

The study of technological networks such as the Internet, has provided important insights into genetic and metabolic networks. The methods of comparison are qualitative, rather than quantitative. The currently known network types (scale-free, small-world, modular and random networks) all have been observed in various biological systems [81]. Identification of network type is based on graph-measures, such as clustering coefficients, betweenness centrality, etc. [82,83]. In principle, any numeric measure that can be "locally" computed for vertices or edges of a graph, can be used to draw a distribution. In fact, the main network types are currently qualitatively defined by these distributions. For instance, the number of connections (i.e. the degree) in scale-free networks that are characteristic of many biological systems can be described by a power-law type degree distribution [84]. They can be further subdivided into groups based on the distribution of betweenness-centrality [85], or of the local clustering coefficient[85]. Network patterns defined as small directed subgraphs are also used to characterize network classes, statistical studies revealed similar network patterns shared by genetic and electronic networks [86-88].

4. Conclusions

The framework of entity-relationship networks provides a simple method to describe similarity groups (equivalence classes), patterns, similarity measures that are used in the comparison of sequences as well as protein 3D structures. The strength of this analysis is shown by the fact that it can be extended to the analysis of large and fuzzy structures, such as genomes and networks.

Acknowledgements

This review is partly based on the lectures of the course "Bioinformatics: Computer applications in molecular biology", held in Trieste, Italy, 1992-2003. Special thanks are due to M. Bishop (Hinxton, UK), E. Gasteiger (Geneva, Switzerland), R. Harper (Hinxton, UK). D. Judge (Cambridge, UK), D. Landsman (Bethesda, MD), J. Leunissen (Wageningen, The Netherlands) for advice, as well as the following individuals for their comments on various topics in the manuscript: Stephen Altschul, Steve Bryant (Bethesda, US), Alexandre De Leon, (Calgary, Canada) Jacques Demongeot (Grenoble, France), Mark Gerstein (Newhaven, CT, UK), Andrew Harrison (London, UK), Lisa Holm (Hinxton, UK), Christine Orengo (London, UK) and William F. Pearson (US)

References

[1] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48 (3):443-53.

[2] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85 (8):2444-8.
[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215 (3):403-10.
[4] Koonin EV, Galperin MY. Sequence, evolution, function. Boston, Dordrecht, London: Kluver Academic Publishers, 2003.
[5] Higgins D, Taylor WR. Bioinformatics, Sequence, structure, and databanks. Oxford, New York: Oxford University Press, 2000.
[6] Smith TF, Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981;147:195-7.
[7] Gumbel EJ. Statistics of extremes. New York, NY.: Columbia University Press, 1958.
[8] Karlin S, Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci U S A 1993;90 (12):5873-7.
[9] Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 1990;87 (6):2264-8.
[10] Krause A, Vingron M. A set-theoretic approach to database searching and clustering. Bioinformatics 1998;14 (5):430-8.
[11] Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set. Nucleic Acids Res 2000;28 (1):270-2.
[12] Attwood TK. The role of pattern databases in sequence analysis. Brief Bioinform 2000;1 (1):45-59.
[13] Murvai J, Vlahovicek K, Pongor S. A memory-based approach to protein sequence similarity searching. In: Pifat G, editor. Supramolecular Structure and Function. Dordrecht/Plenum Press, New York, USA: Kluwer Scientific, 2001. pp. 167-84.
[14] Murvai J, Vlahovicek K, Pongor S. A simple probabilistic scoring method for protein domain identification. Bioinformatics 2000;16 (12):1155-6.
[15] Stanfill C, Waltz D. Toward memory-based reasoning. Communications of the ACM 1986;29 (12):1213-28.
[16] Murvai J, Vlahovicek K, Szepesvari C, Pongor S. Prediction of protein functional domains from sequences using artificial neural networks. Genome Res 2001;11 (8):1410-7.
[17] Vlahovicek K, Carugo O, Murvai J, Pongor S. Prediction of protein structure and function: Towards a memory-based interpretation of proteome data. In: Gromiha M, Selvaraj S, editors. Recent Research Developments in Protein Folding Stability and Design. Trivandrum, India: Research Signpost, 2001. pp. 141-50.
[18] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29 (1):22-8.
[19] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25 (17):3389-402.
[20] Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST- constructed position-specific score matrices. Bioinformatics 1999;15 (12):1000-11.
[21] Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29 (14):2994-3005.
[22] Hobohm U, Sander C. A sequence property approach to searching protein databases. J Mol Biol 1995;251 (3):390-9.
[23] Werner T. The promoter connection. Nat Genet 2001;29 (2):105-6.
[24] Werner T. Promoter analysis. Ernst Schering Res Found Workshop 2002;38:65-82.
[25] Frech K, Danescu-Mayer J, Werner T. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol 1997;270 (5):674-87.
[26] Solovyev VV, Makarova KS. A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. Comput Appl Biosci 1993;9 (1):17-24.
[27] Solovyev VV, Salamov AA. INFOGENE: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. Nucleic Acids Res 1999;27 (1):248-50.
[28] Pongor S. The use of structural profiles and parametric sequence comparison in the rational design of polypeptides. Methods Enzymol 1987;154:450-73.
[29] Remington SJ, Matthews BW. A systematic approach to the comparison of protein structures. J Mol Biol 1980;140 (1):77-99.

[30] Remington SJ, Matthews BW. A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. Proc Natl Acad Sci U S A 1978;75 (5):2180-4.
[31] Levine M, Stuart D, Williams J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. Acta Cryst. 1984;A40:600-10.
[32] Karpen ME, de Haseth PL, Neet KE. Comparing short protein substructures by a method based on backbone torsion angles. Proteins 1989;6 (2):155-67.
[33] Rackovsky S. Quantitative organization of the known protein x-ray structures. I. Methods and short-length-scale results. Proteins 1990;7 (4):378-402.
[34] Rackovsky S, Scheraga HA. Influence of ordered backbone structure on protein folding. A study of some simple models. Macromolecules 1978;11 (1):1-8.
[35] Orengo CA. A review of methods for protein structure comparison. In: Taylor WR, editor. Patterns in Protein Sequence and Structure. Heidelberg: Springer-Verlag, 1992. pp. 159-88.
[36] De Brevern AG, Benros C, Gautier R, Valadie H, Hazout S, Etchebest C. Local backbone structure prediction of proteins. In Silico Biol 2004;4 (2):0031.
[37] de Brevern AG, Valadie H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 2002;11 (12):2871-86.
[38] Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. J Mol Biol 2004;339 (3):591-605.
[39] Guyon F, Camproux AC, Hochez J, Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic Acids Res 2004;32 (Web Server issue):W545-8.
[40] Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208 (1):1-22.
[41] Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. Proteins 1992;14 (2):139-67.
[42] Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233 (1):123-38.
[43] Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A 1998;95 (11):5913-20.
[44] Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 1993;6 (4):377-82.
[45] Andersen CA, Bohr H, Brunak S. Protein secondary structure: category assignment and predictability. FEBS Lett 2001;507 (1):6-10.
[46] Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure (Camb) 2002;10 (2):175-84.
[47] Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. J Mol Biol 2000;301 (3):665-78.
[48] Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 2000;301 (3):679-89.
[49] Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J Mol Biol 2000;301 (3):691-711.
[50] Alexandrov NN, Fischer D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins 1996;25 (3):354-65.
[51] Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3 (2):71-84.
[52] Mizuguchi K, Go N. Comparison of spatial arrangements of secondary structural elements in proteins. Protein Eng 1995;8 (4):353-62.
[53] Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23 (3):356-69.
[54] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6 (3):377-85.
[55] Martin AC. The ups and downs of protein topology; rapid comparison of protein structure. Protein Eng 2000;13 (12):829-37.
[56] Murthy MR. A fast method of comparing protein structures. FEBS Lett 1984;168 (1):97-102.
[57] Mitchell EM, Artymiuk PJ, Rice DW, Willett P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 1990;212 (1):151-66.
[58] 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 (3):707-21.

[59] Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. J Mol Biol 2002;323 (5):909-26.
[60] Carugo O, Pomgor S. Protein fold similarity estimated by a probabilistic approach based on Calpha-Calpha distance comparison. J. Mol. Biol. 2002;315 (4):887-98.
[61] Rogen P, Fain B. Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci U S A 2003;100 (1):119-24.
[62] Rogen P, Bohr H. A new family of global protein shape descriptors. Mathematical Biosciences 2003;182 (2):167-81.
[63] Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet 2002;18 (9):472-9.
[64] Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 2003;3 (1):2.
[65] Hegyi H, Lin J, Greenbaum D, Gerstein M. Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. Proteins 2002;47 (2):126-41.
[66] Gerstein M, Hegyi H. Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 1998;22 (4):277-304.
[67] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000;407 (6804):651-4.
[68] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001;411 (6833):41-2.
[69] Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr., Kyrpides N, Fonstein M, Maltsev N, Selkov E. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000;28 (1):123-5.
[70] Podani J, Oltvai ZN, Jeong H, Tombor B, Barabasi AL, Szathmary E. Comparable system-level organization of Archaea and Eukaryotes. Nat Genet 2001;29 (1):54-6.
[71] Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. Genome Res 1999;9 (6):550-7.
[72] Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet 1999;21 (1):108-10.
[73] Korbel JO, Snel B, Huynen MA, Bork P. SHOT: a web server for the construction of genome phylogenies. Trends Genet 2002;18 (3):158-62.
[74] Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000;28 (18):3442-4.
[75] Huynen MA, Snel B. Gene and context: integrative approaches to genome analysis. Adv Protein Chem 2000;54:345-79.
[76] Huynen M, Snel B, Lathe W, Bork P. Exploitation of gene context. Curr Opin Struct Biol 2000;10 (3):366-70.
[77] Lathe WC, 3rd, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends Biochem Sci 2000;25 (10):474-9.
[78] Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 2002;30 (10):2212-23.
[79] Kanehisa M. Post-genome informatics. Oxford New York: Oxford University Press, 2000.
[80] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28 (1):27-30.
[81] Oltvai ZN, Barabasi AL. Systems biology. Life's complexity pyramid. Science 2002;298 (5594):763-4.
[82] Bollobás B. Random Graphs. Cambridge: Cambridge University Press, 2001.
[83] Dorogovtsev SN, Mendes JFF. Evolution of Networks. Oxford: Oxford University Press, 2003.
[84] Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999;286 (5439):509-12.
[85] Goh KI, Oh E, Jeong H, Kahng B, Kim D. Classification of scale-free networks. Proc Natl Acad Sci U S A 2002;99 (20):12583-8.
[86] Yook SH, Jeong H, Barabasi AL. Modeling the Internet's large-scale topology. Proc Natl Acad Sci U S A 2002;99 (21):13382-6.
[87] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science 2002;298 (5594):824-7.
[88] Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 2002;31 (1):64-8.