Essays in Bioinformatics
D.S. Moss et al. (Eds.)
IOS Press, 2005

Concepts of Similarity in Bioinformatics

Vilmos ÁGOSTON[1], László KAJÁN[2], Oliviero CARUGO[2,3], Zoltán HEGEDÜS[1], Kristian VLAHOVICEK[2,4] and Sándor PONGOR[2]
[1]*Bioinformatics Group, Biological Research Center, Hungarian Academy of Sciences, Temesvári krt. 62, 6726 Szeged, Hungary*
[2]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy*
[3]*Department of General Chemistry, Pavia University, viale Taramelli 12, 27100 Pavia, Italy*
[4]*Molecular Biology Department, Biology Division, Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia*

Abstract. The key problem of bioinformatics is the prediction of properties, such as structure or function, based on similarity This chapter reviews the concepts and tools of similarity analysis used in various fields of bioinformatics.

Introduction

The concept of similarity is fundamental in the study of macromolecular structures, genomes, proteomes and metabolic pathways. Similar objects are often assumed to take part in similar mechanism, or to carry out a similar function. Similarity, on the other hand is a highly intuitive concept, and its use in various fields – such as the comparison of sequences or of 3-D structures – is quite different. For students of molecular biology it is sometimes difficult to find straightforward definitions of the basic concepts that originate from as diverse fields as cognitive psychology, systems science as well as various branches of mathematics. The motivation of this review is to provide a – not necessarily complete - compendium of useful concepts and definitions and to show the commonalities underlying the various applications. We will use three main forms of representations: sequences, 3-D structures and graphs. The discussion will be based on an entity-relationship description of macromolecular structures [1], as applied to the description of small molecules [2] as well as biological objects used in genome analysis [3].

Most concepts of molecular similarity have been proposed in applied contexts that are so numerous that an exhaustive coverage would detract from our focus on the underlying mathematical spaces. In particular, machine learning methodologies used in bioinformatics [4, 5], such as neural networks [6] and support vector machines [7] are based on specific concepts that in our view cannot be adequately described in the framework of a general discussion. Similarly, we could not include a practice-oriented overview of applications such as the comparison of sequences, 3D structures and genomes (a review on these topics will be published elsewhere [8]). Several fields that are gaining importance in bioinformatics, such as the analysis text similarities [9], could not be incorporated because of space limitations. Although a significant amount of research is thus excluded from this overview, a broad, and we hope to show, integrated body of research remains.

The primary focus of this work is to present a set of useful definitions pertinent to the similarity analysis of macromolecular structures, meant as reference material for

advanced bioinformatics courses. *Section 2* describes the basic concepts used in macromolecular similarity analysis, pointing out, whenever possible, the parallel concepts in other fields. *Section 3* focuses on four distinct mathematical relationships, each of which constitutes a possible definition of similarity: equivalence, matching, partial ordering, and proximity.

1. Basic concepts

1.1 Model, description, analysis

When we speak about molecules, what we mean are not physical entities, rather abstract models of reality. It is useful to distinguish three concepts underlying molecular data:

The *models* are the conceptual structures or mental representations used to store information on molecules. These models never incorporate all of information available on a given macromolecule – the mere listing of the atoms and bonds in a macromolecule would be beyond the reach of human memory – rather we deal with a set of models of varying complexity, each describing a certain aspect of the molecular structure, such as linear sequence, domain topology, active site contacts, etc.

Various formal and/or narrative *descriptions* of the data constitute the backbone of molecular databases. We can imagine the descriptions as the mathematical representation of a particular model. Similarity measures are calculated between descriptions (and not between models).

The *analysis* covers everything we do with molecular data in such fields as molecular modelling, prediction, classification, similarity search, visualization etc.

For example we may start noticing a new regularity when classifying the existing molecular descriptions (*analysis*). If this new feature "makes sense" (e.g. it points to a meaningful subclass of the objects) we may include this into our abstract *model*, and we may proceed to construct a new kind of description that includes the new feature. In a further round of analysis we may find new examples that contain the feature in question, in addition we may experiment with new feature candidates analogous or similar to the previously found features. As this cycle is repeated, the models and the descriptions undergo an evolutionary change, and in fact this is how databases develop [10].

1.2 Entities, relationships, structure and function

In the first approximation, bioinformatics is concerned with the structure of protein and DNA molecules that fulfil functions in a series of interdependent systems such as pathways, cells, tissues, organs and organisms. This complex scenario can be best described with the concepts of systems theory (**Figure 1**).

According to systems theory [11, 12], a system is a group of interacting elements functioning as a whole and distinguishable from its environment by recognizable boundaries Molecules can be regarded as such systems. Generally speaking, *structure* is fixed state of a system, and the study of a system usually starts with its characteristic structures that are recurrent in space or time. As structures are detected by recurrence, the symmetries (internal repetitions) are integral parts of structural descriptions. Using the terms of the previous paragraphs, systems are conceptual models of reality, while structures are descriptions.

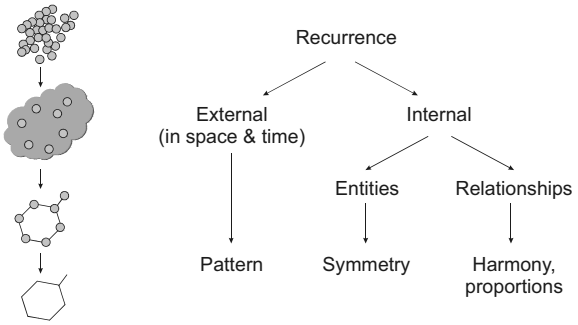

Figure 1. Simplified overview of concepts underlying structural descriptions.

Descriptions rely on elements (entities) and binary relationships between them [1, 13] (Table 1).

In the case of molecules, both the elements (substructures) and the relationships can be described in terms of systems of categories. The categories and the relations between them can be formalized into ontologies, which include the definitions of the elements as well as the operations that are possible within the system (**Figure 2**). Ontologies give itemized descriptions each functions and roles a molecule can fulfil, so it is a logically coherent world description. Entity-relationship-descriptions are generally applicable and can be extended to such concepts as similarity groups, vicinities and networks (**Figure 2**).

Table 1. Examples of models and descriptions

System	Entities	Relationships
a) Conceptual models of natural systems		
Molecules	Atoms	Atomic interactions (chemical bonds)
Assemblies	Proteins, DNA	Molecular contacts
Pathways	Enzymes	Chemical reactions (substrates/products)
Genetic networks	Genes	Co-regulation
b) Structural descriptions		
Protein structure	Atoms	Chemical bonds
Protein structure	Secondary structures	Sequential and topological vicinity
Folds	C_α atoms	Peptide bond
Protein sequence	Amino acid	Sequential vicinity

Elements and relationships can be described not only in terms of categories, but we can assign to them property descriptors, such as physicochemical, chemical descriptors. In terms of contents, there are two kinds of properties in proteins and DNA that deserve special attention. i) The *position* of an element (nucleotide, atom) can be defined either within the molecular chain (sequential position, with respect to the N-terminus, etc.) or in as 3-D coordinates. ii) The *function* is a property or role that can be defined in the context of a higher level. E.g. “protease” is a function defined either in an *in vitro* (e.g. action on a certain substrate) or *in vivo* environment (e.g. role in complement activation). In addition to these two main classes, there are a whole list of properties that can be assigned to entities

and relationships within a model. In terms of mathematical form, the descriptors of the properties can be continuous, discrete or binary variables, even statements in human language.

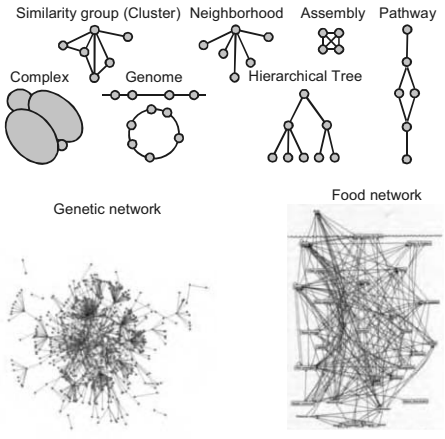

Figure 2. Molecular structures can be represented as entities and relationships [1, 13]. Implicit to a structure is the description of the underlying concepts (entities and relationships as well as their properties), which can be summarized in an ontology [14]. The same principle can be easily extended to genomic and "systems biology" applications.

Entity/relationship models have been used in psychology as well. Erich Goldmeier's "Similarity of visually perceived forms" defines similarity in terms of partial identities that may include a varying proportion of entities and relationships [15, 16]. If we apply this definition to molecular graphs such as shown in Figure 2, we arrive to a plausible definition: Two molecular graphs are similar if they have a common sub-graph (Figure 3).

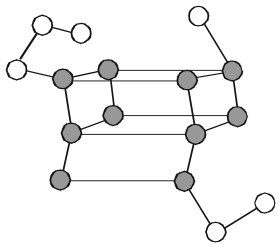

Figure 3. Molecular similarity as sub-graph isomorphism. Similarity of structures can be defined as a common sub-graph shared by two entity-relationship descriptions.

Dedré Gentner [17] drew a map classifying the similarities of narrative descriptions (Figure 4a), which can be extended without difficulties to the description of protein structures (Figure 4b). For example, molecular descriptions are considered identical if they consist of the same substructures and relationships. If two descriptions only share the substructures but not the relationship, they are identical in terms of composition only. If the relationships are identical, but not the substructures, we speak about equivalent topology. Alpha-helices (and other protein secondary structure elements) are examples for this kind

of partial identity, since in this case the identity of amino acid residues (i.e. the entities) is immaterial. All identities and similarities are true only at the given level of description (e.g. backbone conformation, amino acid composition, etc.).

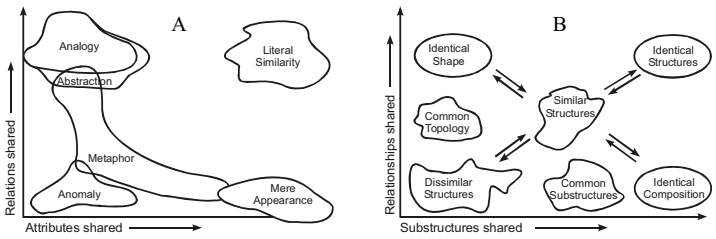

Figure 4. Identity, different kinds of similarity and non/identity can be pictured as regions in a plot of shared entities vs. shared relationships. This representation was developed by Dedré Gentner for narrative descriptions [17] (A), but can be extended to molecular descriptions as well (B).

Figures 4 implies that similarity of two molecules can be captured if we can define equivalencies between their constituents, i.e. if we match the similar parts of the two descriptions to each other. Finding common substructures relies on matching, and some numerical parameter of matching is used in most cases as a measure of similarity. For example, two 3D structures are obviously similar if more than 90% of their alpha carbons can be superposed. We mention that matching is used not only for establishing similarity, but also for finding complementarity, such as surface-complementarity used in molecular docking, or strand-complementarity used in the analysis of anti-sense RNA.

Based on the above concept we can define two further concepts, similarity groups and functional units. The similarity group is such a group of molecules that are connected by structural similarity. This similarity can be local or global (see 2.3) or it can be general or specific (section 3.3). Biologically important similarity groups, such as those of protein domains belong to the latter class, as all group members are characterised by a common sequence-description or a common fold-description.

Functional units denote a group of molecule that jointly fulfil a biological function. Enzymes, regulators and substrates of a metabolic pathway are examples of functional unit. Members of a functional unit are similar in their common function, but they do not need to be structurally similar. This is thus a contextual similarity, as opposed to the structural similarity.

1.3 Elements of molecular descriptions

1.3.1 Focusing of descriptions

The entity-relationship framework and the underlying category definitions can be used to construct a very large number of description that can focus on various aspects of a molecular model [13]. One of the practical ways of generating simplified descriptions is to concentrate on parts of a molecule that are important for the actual goal of the analysis. Starting from a generalized theoretical model containing detailed descriptions of all entities and relationships in various forms, one can derive simplified descriptions by omitting some of the descriptors. For example, a hydrophobicity plot is a description of protein structure wherein the entities are amino acid residues described in terms of only two parameters, the sequence position and the residue hydrophobicity index. On the other hand fold

descriptions include only the C_α atoms of a protein, while surface descriptions include only those atoms in contact with the environment (solvent). But we may choose to use higher categories, such as domain-units instead or amino acid residues. TOPS cartoons are simplified description in which the entities are secondary structural elements; the relationships are topological links describing sequential or spatial vicinities.

Table 2. An example of simplified descriptions

Model	Descriptor	
	Position	Hydrophobicity
Hydrophobicity plot	+	Real number
Hydrophobic segments	+	Discrete (0 or 1)
Average hydrophobicity	-	Real number
Hydrophobic character	-	"Hydrophobic"/"Hydrophilic"

Another avenue of fine-tuning consists in decreasing the detail - the resolution - of the descriptors (Table 2). For example, residue hydrophobicity can be described in quantitative terms, using a hydrophobicity scale (with continuous variable represented as a real number) or qualitatively (discrete variable, represented as 0 or 1 or with categories "hydrophobic" and "hydrophilic").

The intuitive concept of *resolution* also refers to the number of categories used in a given description. An amino acid composition is a vector in a 20-dimensional space, and since most proteins contain all of the amino acids, all the components of the vector are non-zero. On the other hand, we have 400 dipeptides and 8000 tripeptides. In a tripeptide-based composition, however, many (or most) of the components would be zero or 1. Very high-resolution descriptions are highly characteristic "fingerprints" that can be used to identify individual structures. For example, mass spectra are efficiently identified by the presence/absence of their constituent peaks, and similarly, small molecular structures can be retrieved from databases using queries constructed from their constituent fragments. On the other hand, high-resolution fingerprints cannot be easily generalized to similar molecules, so the resolution of the descriptions has to be optimized so as to include the right scope of similar descriptions.

1.3.2 Kinds of descriptors

Descriptors can be categorized according to their contents. On the one hand we have various levels, such as atoms, residues, secondary structure element, domain etc. Whether we talk about DNA or about proteins, there is an apparent lowest level that is not divided into further categories. For example, structural biology is rarely concerned with particles below the atomic level, while molecular biologists use nucleotides and amino acids as the lowest level. Higher-order units can be built up from the lower levels. In most cases the higher units are non/overlapping, i.e. one atom can be part only with one residue. On the other hand we use overlapping fragment descriptions as well, for example nucleotide sequences can be described in terms of overlapping di- or trinucleotide words, protein 3D structures can be described as peptide fragments.

We use the term "*structured descriptions*" for those descriptions that contain both entities and relationships. Protein 3-D structures and sequences are such descriptions even though the relationships are not explicitly included in the actual descriptions found in

databases. For example, the atoms are named in PDB files, but the connectivity of atoms in amino acids is not part of the database, it rather has to be included in the program reading the database entries. If a description contains only entities or only relationships, we term it an "*unstructured description*". Examples include amino acid composition (only entities) and C_α distance-distributions (only relationships).

Finally, descriptors can be classified also depending on what they refer to. Descriptors referring to an entire molecule are global descriptors, such as a protein function. Local descriptors, such as the role of a domain within the protein are local descriptors.

1.4 Overview of macromolecular descriptions

Based on the concepts introduced in the preceding sections we can now attempt to classify the molecular descriptions. One simple classification distinguishes 1D, 2D and 3D descriptions. 1D descriptions, such as sequences and hydrophobicity plots, are residue-based, and include only the chain-topology. 2D descriptions are graph-like and include relations in addition to the chain topology (e.g. helical circle and helical net diagrams provide a symbolic view of the 3D arrangements). 3D descriptions are those in which Cartesian coordinates are included among the descriptors.

A more detailed classification is possible according to the mathematical machinery. This classification essentially follows that of Johnson set up for small molecules [2, 18].

The most complete description is a generalized labelled graph in which both the vertices, and the edges can be provided with arbitrary labels such as numbers, vectors, names even statements in human language. Labels can be attached to individual entities or to groups of them (such as segments of a polypeptide chain). This is a hypothetical, multi-level description that is best approximated by a well-annotated 3D database record that is cross-referenced to (possibly all) the available biological databases. Such variable-level descriptions are rarely used for comparison. The 3D comparison programs of Sali and Blundell are one of the few exceptions, they use a hierarchy of levels such as atoms, residues, secondary structures and domains [19, 20].

3D structures contain atoms and entities provided with Cartesian coordinates as descriptors, as well a chemical (covalent) connectivity. This description is used by most of the molecular modelling and structure comparison programs. Structural databases contain the entities and their labels; the connectivity maps are included with the analysis programs.

Distance matrices. Distances calculated between the elements of the same structure constitute a distance matrix. In 3D structures, one can use the positional coordinates to define distance vectors, whereas the number of edges between two nodes can be used to define a distance in a graph. Both are extensively used in similarity analysis.

Finite sequences. All graphs can be represented in terms of finite sequences. A protein sequence is a special graph where the residues are the entities and the polypeptide chain connectivities are the edges. 1D plots (such as the hydrophobicity plot) can be derived from an amino acid sequence by representing one single numeric parameter as a function of the residue position. This parameter can be either an experimentally determined value (such as a physicochemical parameter, or a quantity computed from the sequence or from the 3D structure.

Surfaces used for proteins include the Van der Waals surface or the electrostatic surfaces that are computable from the 3D structure. Surface similarity analysis is not included in this review, an excellent review is in [21-23].

Integrable scalar fields. In this representation the molecule is treated as a spatial distribution of a single quantity, such as electron density or mass density [24].

Transforms. There are various methods to calculate topological transforms from graphs. Fourier transforms of 1D sequence plots have been used to identify amphifilic regions in proteins, as well as to compare proteins.

Finite sets are unstructured descriptors that can be obtained e.g. by omitting all relationships from a labeled graph. The resulting set of entities provides description that can be ordered according to kinds. A typical example is the amino acid composition, or other fragment-composition type descriptions (dipeptide, tripeptide etc. compositions). This is a vector-representation, the parameters of the vector corresponds to the number of times a certain entity is present in a structure A subcase of finite set descriptions consists in reducing the set of entities to a set (list) of kinds. This can be achieved by omitting the numbers from a compositional description.

Distributions. A vector consisting of nonnegative numbers that sum to unity constitutes a parameter vector of a multinomial distribution. A typical example is the amino acid composition expressed in percentages, or the distribution of inter-atomic distances within a protein structure, or distribution of connectivity degrees in large networks.

Vectors, product spaces. In addition to the special vectors mentioned in 5 and 6, arbitrary parameters of a given molecules can be assembled into vectorial descriptions. Such complex descriptions are used as input in machine-learning, and are also often used in general pattern-recognition applications.

Real numbers (molecular sizes, molecular weight etc.) are perhaps the simplest descriptors of molecules.

2. Mathematical concepts related to similarity

2.1 Relations

2.1.1 Equivalence

Equivalence relations (denoted here by "≅") are related to the commonly used term of identity. Strictly speaking, a molecule can only be identical with itself; here we are concerned with the cases when two molecules have identical mathematical descriptions, which does not mean that they are identical. For example, two proteins that have an identical description in terms of amino acid sequence may undergo phosphorylation or other posttranslational modifications at different sequence positions).

Equivalence relations in mathematics are defined by three properties: reflexivity, symmetry, and transitivity. A relation is reflexive if $A \cong A$ for all molecular descriptions A. It is symmetric if $A \cong B$ implies $B \cong A$. It is transitive if $A \cong B$, and $B \cong C$ implies $A \cong C$. Let [A] denote the family of those molecules equivalent to A with respect to ≅. If B denotes some other molecule, it can be proven mathematically that either [A] and [B] denote the same set of molecules or the two sets have no members in common. The set [A] is called an equivalence class. For example two proteins are considered identical if and only if their (amino-acid) sequences are the same. It is noted that "identity" refers to a given description; in this example the potential differences in post-translational modifications are disregarded.

2.1.2 Partial ordering

Partial ordering relations are related to the commonly used terms "to be a substructure of", "to be a part of". A relation ≤ is called a partial order if it is reflexive, antisymmetric, and transitive. The reflexive and transitive properties of a relation were defined earlier. A

relation is antisymmetric if A ≤ B and B ≤ A implies that A and B are identical. For example if A ≤ B means that A is a subsequence of B, then ≤ is a partial ordering relation.

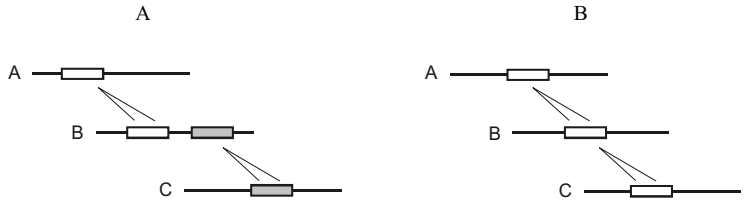

Figure 5. Similarity of molecules can be considered either a tolerance relationship (A), or an equivalence relationship (B) depending on whether or not the basis of similarity – the shared substructure – is fixed.

2.1.3 Tolerance, general and specific similarity

Tolerance relations denote the common sense situation in which two things have a common part or feature, or two structures share a common substructure. A relation ~ is called a tolerance if it is reflexive, symmetrical, but – in contrast to equivalence relations – not necessarily transitive. In other words, A~A, A~B implies B~A. Tolerance comes closest to the common sense concept of similarity, however there is an important distinction to be made. Based on the psychological concept of Goldmeier [15, 16], we can call two structures similar if they share some common substructure (see **Figure 3**, above). This general similarity is not transitive, as shown in **Figure 5a**, it is in fact a tolerance relationship. On the contrary, we may use the term specific similarity, if two structures share a well-defined substructure (feature). Fixing the shared substructure renders the relationship transitive, so specific similarity is an equivalence relationship (**Figure 5b**).

If biological sequences are found similar to each other by BLAST, this is a general similarity, i.e. it is not necessarily true that all of them share a subsequence, such as a protein domain. However, those sequences that turn out to share a common subsequence form an equivalence class. It is noted that a "common subsequence" is often defined in an empirical way: biologists usually decide based on their prior knowledge whether or not a subsequence of a protein is a true member of a domain group (like EGF domains), and once a positive decision is made, the protein sequence is accepted as a member of the equivalence class of EGF-containing proteins. We might say that evaluation of BLAST searches consists in distinguishing general and specific similarity.

The use of relations in chemical structure analysis is reviewed in [2, 18].

2.2 Proximity measures

Proximity measures (*PM*) are numeric measures designed to characterize similarity or dissimilarity of two molecular descriptions. Two general types of proximity measures are in use. Similarity measures are high for similar molecules and low for dissimilar ones. The distance measures, on the other hand are zero for identical molecular descriptions and high for dissimilar ones. In the foregoing we will use proximity measures, distance measures and similarity measures.

Proximity measures can be used in vastly different contexts, and it is useful to define two situations that are common in bioinformatics applications. A) *Simple proximity*

between two objects is computable by an unequivocal algorithm. A distance of two stars in space is a good example. For instance, such measures are computed between unstructured descriptions like vectors. One can define simple proximity measures also for structured descriptions, provided the equivalences of the entities (residues, atoms) are *a priori* defined. The Hamming distance and *rmsd* are such measures for character strings and 3D structures respectively. They are based on straightforward algorithms for calculating a distance between the two objects. Such distances are often calculated between fragments of larger structures hence they are sometimes called fragment distances. B) *Substructure proximity* measures are computed between parts structured descriptions. They require a simple measure as well as an algorithm to select the "optimal substructure" in the two objects. For instance, the distance of two galaxies can be defined as the distance between their closest stars. In this case, we have to measure the distance between all possible pairs (simple proximity), and then select the smallest one. The two central problems of bioinformatics, – sequence alignment and 3D structural alignment – are substructure similarity problems. So instead of two objects, we need compare "galaxies of substructures" which is a compute intensive task. The complexity of the calculation is different (sequence alignment has a complexity *O(n,m)* while structural alignment is *np*-complete), but the the basic concepts of substructure selection matching– described in the subsequent paragraph 3.3 – is common to both. The present section will concentrate mostly on simple proximity measures. We will follow the classification of Sneath and Sokal [25] who distinguished four classes proximity measures: distance coefficients, association coefficients, correlation coefficients and probabilistic coefficients.

2.3 Distance measures

Vector distance measures are perhaps the simplest class of similarity measures owing to their geometric interpretation. The most common is probably the Euclidean distance, which, for some pair of objects A and B, described by n-dimensional vectors A_i and B_i, respectively, is defined as:

[1] $$D_E = \left(\sum_{i=1}^{n} (A_i - B_i)^2 \right)^{\frac{1}{2}}$$

This is a distance defined in the n-dimensional space. A simple variant of this formula is the average distance, which is simply the Euclidean distance divided by the number of dimensions, i.e., by *n* in this case. The generalization of the Euclidean distance leads to a class of metric distance functions called the Minkowski metrics, defined by the following general formula:

[2] $$D_M(r) = \left(\sum_{i=1}^{n} |A_i - B_i|^r \right)^{\frac{1}{r}}$$

$r = 1$ corresponds to the "city block" distance and $r = 2$ to the Euclidean distance. The so-called sup distance is the Minkowski metric of $r = \infty$ and corresponds to

[3] $$D_M(\infty) = \max_{1 \le i \le n} |A_i - B_i|$$

Distance measures calculated between identical molecular descriptions (vectors) are zero, and may grow without limit for non-identical vectors. It is sometimes desirable to have bounded values, for example the so-called Canberra metric is defined as:

[4] $$C_M = \frac{\sum_{i=1}^{n} |A_i - B_i|}{\sum_{i=1}^{n} (A_i + B_i)}$$

so it is zero for identical values but remains below unity for non-identical vectors. The metric properties of vector distance measures are important for clustering and for evolutionary studies. For M to be a metric (metric distance), the following criteria have to be fulfilled for all A, B, C from X: *i)* $M(A,B) \geq 0$ the equality holding if and only if $A = B$; *ii)* $M(A,B) = M(B,A)$ (symmetry); *iii)* $M(A,B) + M(B,C) \geq M(A,C)$ (triangular inequality). Metric properties are essential if a distance measure is to be used for clustering. A string similarity measures S (eqn. 5) is applicable to clustering if there is an associated distance measure $M = f(S)$ that has metric properties. f is a monotonous function, and distance measures such as *1-kS* (where k is a constant) are routinely used in clustering applications.

2.4 String similarity measures for biological sequences

A special class of proximity measures, sequence similarity scores are used to quantify the matching (alignment) of protein and DNA sequences. The underlying mathematical concept is the string distance. Let us first concentrate on bit strings consisting of zero/one values. (**Figure 6A**). The Hamming distance is the number of (zero to one or one to zero) changes necessary to change string 1 to string 2. This can be used immediately for short character strings of identical length, with the condition that only exchanges are possible, gaps are not allowed. If we keep this condition, we can use a simple lookup table to store the costs of exchanging one character against another. The situation is the same if we use overlapping doublet or triplet words, etc. i.e. the Hamming distance is a simple distance that can be unequivocally computed based on lookup tables, because the matching of the two strings is considered unique.

The situation is quite different if we match strings of arbitrary length and allow gaps (Figure 7). The string edit distance is defined as the minimal number of steps (insertions, deletions and replacements) necessary to transform one word into the other. The proximity measures used for biological sequences are defined as similarity coefficients (high values for similar, low for dissimilar sequences) and contain cost factors for residue substitutions as well as gaps (insertions, deletions).

[5] $$S_{1,2} = \sum \mathrm{cost}_{identities,replacementss} - \sum \mathrm{cost}_{gaps}$$

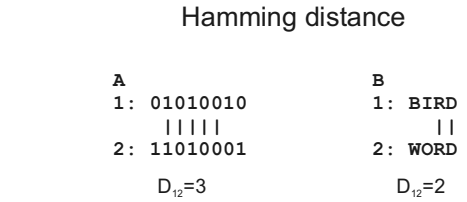

Figure 6. The Hamming distance is the number of exchanges necessary to turn one string of bits or characters into another one (the number of positions not connected with a straight line). It is assumed that the two strings are of identical length and that no alignment is necessary. The exchanges in character strings can have different costs, stored in a lookup table. In this case the value of the Hamming distance will be the sum of costs, rather than the number of the exchanges.

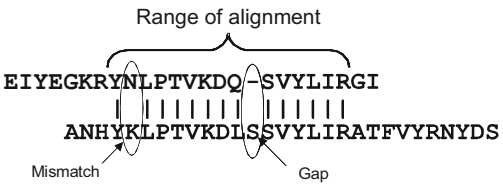

Figure 7. A string similarity measure can be defined as a sum of costs assigned to matches, replacements and gaps (insertions and deletions). The two strings do not need to be of the same length. A string similarity measure between biological sequences is a maximum value calculated within a range of alignment. The maximum depends on the scoring system that includes a lookup table of costs, such as the Dayhoff matrix, and the costing of the gaps.

The alignment used here is no longer unique, like in the case of a Hamming distance, and there are different (arbitrary) ways to cost gaps (different cost factors for gap opening and gap extension etc.). Establishing an alignment between two sequences consists in maximizing a similarity measure given in equation [5]. This problem can be solved if in addition to the formula of *S* we have a cost matrix for replacements and identities, or some other lookup table that contains the similarity/distance values of the elements used in the description. In the case of proteins, the cost factors of amino acid substitutions are included in the well-known Dayhoff and BLOSUM matrices, and there are several established strategies for costing the gaps – for recent reviews see (). The algorithm for finding a maximal similarity between two longer sequences is an optimization problem. The actual algorithms of similarity search are beyond our scope. The basic principle is mentioned in section 3.5, and some examples are given in section x.x. There are number of comprehensive reviews on this subject.

2.5 The rmsd distance for protein 3-D structures

A very popular quantity used to express the structural similarity of 3-D structures is the root-mean-square distance (*rmsd*) calculated between equivalent atoms, defined as

[6]
$$rmsd = \sqrt{\frac{\sum_i d_i^2}{N}}$$

where d is the distance between each of the N pairs of equivalent atoms in two optimally superposed structures. For the calculation of *rmsd* a range of alignment has to be defined within which the matching of atoms (establishment or equivalent atoms within the two structures) is determined which is a computationally much harder problem than the alignment of sequences in one dimension. Once the equivalence of atoms is established, the optimal superposition has to be found which is carried out by such straightforward algorithms as that of Kabsch [26].

If the equivalences are fixed, then rmsd can be considered as a simple distance that can be computed with a straightforward algorithm. This is the case for instance when one compares different conformations of the same protein such as produced by NMR methods. In this case the equivalences of the atoms are *a priori* known, since each conformation consists of the same atoms. The *rmsd* is 0 for identical structures (identical conformations) while its value increases as the two structures become more divergent. In fact *rmsd* values are considered as reliable indicators structural variability when applied to very similar proteins (say *rmsd* < 5-6 A). But even in this case, the *rmsd* value obviously depends on the number of residues N included in the structural alignment. A statistical analysis of a large number of structures showed that the dependence can be described as:

[7] $$rmsd = rmsd_{100}(1 + \ln\sqrt{\frac{N}{100}})$$

where $rmsd_{100}$ is a constant, an *rmsd* value standardized to 100 residues [27]. The *rmsd* values also depend on the crystallographic resolution, which is more difficult to take into consideration (Carugo, 2002). As a result, *rmsd* does not behave as a metric distance for divergent structures so it cannot be used in itself for automated clustering. Clearly, an *rmsd* value of, say 3 Å has a different significance for proteins of 500 residues and for those of 50 residues, so e.g. the structural variability of fold types can not be easily compared ($rmsd_{100}$ on the other hand may be useful for such comparisons[27]). In other terms, *rmsd* is a good indicator for structural identity, but less so for structural divergence.

The algorithms for calculating *rmsd* are beyond our scope, the reader is referred to recent reviews [28]. The philosophy of the calculation depends on whether or not the alignment, i.e. the equivalences between residues (represented as C_α atoms) are known. If yes, the very popular algorithm of [26] and McLachlan (1978) can be used. If this is not the case, and when the two 3-D models that are compared are too different, there are two alternatives. Either a partial alignment is available or no a priori assumptions can be made. In the first case, few equivalences between atom pairs are assumed and they are extended (and some time rejected) through dynamic programming techniques [29]. In the other case an exhaustive search is performed by rotating and translating a 3-D model over the other in a six-dimensional way Diedrichs, 1995).

It has to be noted that superposition of divergent protein 3-D structures is often a quite arbitrary exercise and various superposition algorithms may lead to completely different results. An effective, recently proposed procedure to reconcile different structural alignment procedures consists in an iterative reduction of the number of aligned C_α atom pairs [30]. After each superposition, the worse pair is eliminated and a new superposition is performed leading, eventually, to the identification of the protein core that shows a significant degree of similarity.

Finally we mention that the *rmsd* distance does not allow the costing of gaps. For this reason, it can not be used directly for finding an optimum alignment between two arbitrary proteins.

2.6 Association measures

For the comparison of chemical graphs of small molecules, association measures are used almost as widely as distance measures [31]. The majority of these coefficients are intended for use with simple two-state, i.e., binary, variables which are conventionally coded as 0 or 1 depending upon their presence or absence within an object description. Although these coefficients can be described in terms of a vector it is conceptually simpler to formulate the coefficients as follows. For two objects A and B let ab be the number of attributes in common and $\overline{ab}$ the number of attributes in neither in A nor in B. Let $a\overline{b}$ and $\overline{a}b$ be the number of attributes occurring in only in *A* or *B*, respectively. Let *a* and *b* the total number of attributes in *A* and *B* respectively. Let *n* be the total number of attributes, i.e. *a*+*b*. Two frequently used association coefficients are the Jaccard (also called Tanimoto) coefficient:

[8] $$J = \frac{ab}{a\overline{b} + \overline{a}b + ab}$$

(which can also be written as $a \cap b / a \cup b$, the ratio of common attribute types to all attribute types) and the Dice coefficient

[9] $$D = \frac{2 \cdot ab}{a + b}$$

The coefficients may readily be generalised to non-binary data. For instance, if the data vectors contain the actual frequencies of occurrence of each fragment type, rather than their mere presence or absence, the Jaccard coefficient can be rewritten as

[10] $$J' = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i B_i}$$

A related association measure is the so-called cosine coefficient that corresponds to the cosine of the angle between the vectors *A* and *B*:

[11] $$C = \frac{\sum_{i=1}^{n} A_i B_i}{\left(\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2 \right)^{\frac{1}{2}}}$$

2.7 Correlation measures

Another widely used coefficient of similarity in cluster analysis has been the Pearson product-moment or correlation coefficient. Given two structures *A* and *B*, let $\overline{A}$ be the mean value for all of the variables in the vector *A*(and similarly $\overline{B}$ for). Then the coefficient is defined as

[12]
$$r = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\left(\sum_{i=1}^{n}(A_i - \bar{A})^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}(B_i - \bar{B})^2\right)^{\frac{1}{2}}}$$

The correlation coefficient is 1 for identical vectors, is around zero of dissimilar vectors and is –1 for anticorrelated vectors ($A_i = -B_i$). A large number of related coefficients are given in [32, 33].

2.8 Probability based measures

The final class of coefficients identified by Sneath and Sokal [25], probability based coefficients, take account of the frequency distribution of variables over the entire data set. Probability-based coefficients are less often used for small molecules, on the other hand they are the most often used method of scoring in biological sequence comparison. Probability-based measures are obtained by first calculating a raw proximity measure PM between a query and all members of a dataset. This is followed by rescaling the raw PM using knowledge on the distribution of scores. This operation places the PM values on a common scale and thus provides an obvious way to set significance threshold for the hits of interest. It is customary to distinguish "biologically meaningful" and "random" similarities. The former are those between evolutionarily (homologs, orthologs, paralogs) or structurally related proteins (molecules with a common fold), the rest of the similarities are usually considered "random".

One approach is based on the distribution of random similarities. If the distribution is known in analytical or numeric form, then the statistical significance of any computed measure – the probability P ($0 \leq P \leq 1$) for finding a given value in the given dataset by chance – can be estimated. Random similarities occur more likely for larger queries and for larger databases, so the description of random distributions usually includes query size and database size as variables. (The product *query size x database size* is sometimes referred to as the *search space*). Current biological databases provide a sufficiently large number of data for modeling the distribution of random similarities, and – at least for sequence data – various random shuffling techniques can be used to generate larger datasets. This approach thus consist in rescaling a proximity measure PM to give a probability P for a given search space. This P is called statistical significance, in other words, if the value of proximity measure lies far outside the distribution of random scores (P is very small), one tends to consider it biologically significant, and conversely, large P values indicate random similarities that are unimportant in the biological sense.

Another approach relies on the distribution of the target similarities, i.e. the distribution of PM within a biologically important group of objects. Often there are not enough reliable data for the analytical modelling of this target distribution, and random shuffling techniques may not be easily applicable same as for random similarities. A compromise solution consists in concentrating on the distribution of biologically significant as well as random similarities in the neighbourhood of a target group[34, 35]. This approach relies on the fact that space defined by existing macromolecules is sparsely and unevenly populated (as compared to the hypothetical space of all possible molecules), and the neighbourhoods of existing similarity groups may be quite different.

Further kinds of probabilistic coefficients can be obtained if one represents the objects themselves by some kind of a distribution, and then compares two distributions so as to obtain a probabilistic estimate of their identity [36]. There are established methods for

the comparison of distributions, such as the χ^2 test and contingency table analysis, etc. [37] that all yield probability values between 0 and 1.

Probability-based measures are widely used for the evaluation of prediction methods [32, 33]. Similarity measures for chemical structures have been reviewed by Willett [31].

2.9 Proximity measures for groups of objects

Proximity measures originally defined to pairs of structural descriptions can be generalized to groups. Given a single description *S* and a group of descriptions $[A]=\{A_1, A_2, \ldots A_n\}$, a proximity measure $P(X,Y)$ between *S* and *[A]* can be defined using the $P(S,A_i)$ values of the pairwise comparisons; for example, one can take the minimal, the maximal or the average of the $P(S,A_i)$ values as the proximity measure between *S* and the group. Another possibility is to calculate from the descriptions A_i a "consensus value" $<A>$, sometimes called the centroid of *[A]*. If the descriptions are simple numeric values or vectors, $<A>$ can be defined as their average. If A_i-s are vectors, $<A>$ can be their vectorial average, etc. Then, the proximity measure between *S* and *A* can be calculated as $P(S,<A>)$.

Proximity measures between two groups of objects *[A]* and *[B]* can be defined in a similar way: we can take the minimum, maximum or average of the $P(A_i,B_j)$ proximity measures, or determine the proximity of the two centroids, $P(<A>,<B>)$.

If a single object is compared to group *[A]* in terms of a feature *f* that is supposed to be normally distributed in *[A]*, with mean *m* and standard deviation *sd*, then, instead of the simple difference $|f-m|$ we can use a scaled value $\left|\frac{f-m}{sd}\right|$ for calculating a distance between an object and the group. Similarly, one can calculate a distance between two groups (denoted by upper indices 1 and 2, respectively) using the values $\left|\frac{m^1-m^2}{\sqrt{(sd^1)^2+(sd^2)^2}}\right|$. The resulting distance values will thus incorporate a natural scaling based on the different variance of the groups. This scaling can be generalized to cases in which the objects to be compared are represented as vectors of features $f_1, f_2 \ldots f_n$ characterized by a covariance matrix *C*. In this case, the so-called Mahalanobis distance is defined as:

[13] $$MD = (m^1 - m^2)' C\hat{} (m^1 - m^2)$$

where m^1 and m^2 are average vectors for group 1 and group 2, respectively, $(m^1 - m^2)'$ is the transpose of $(m^1 - m^2)$ and $C\hat{}$ is the inverse of the variance-covariance matrix *C*. *MD* can be viewed as an Euclidean distance scaled by the covariance matrix, the latter being assumed to be identical for both groups.

3. Matching (alignment)

For two structures to be similar, one has to find a matching in terms of entities and relationships. Such a matching is shown in **Figure 3**. A matching resembles an analogy. In

an analogy, features of one object are paired with features of another object. Which features are paired is often a subjective choice. Matching of molecular structure relies in comparing molecular descriptions finding an optimal match between the features of the mathematical representations.

The simplest example of matching is the alignment of short character strings of equal length described in Figure 6. Another example is to find the exact occurrence of a short character string query within another, longer string.

With these examples one can illustrate, without formal definition, some of the important properties of matching used in bioinformatics. Given two descriptions A, B consisting of i and j elements, respectively, let's define a mapping m: $A \rightarrow B$ that assigns certain k elements in A to certain l elements in B; it is not necessary that all elements in A have an element assigned in B and vice versa. In the simplest case, the mapping is one-to-one, so certain elements in A will have a pair in B (so $k=l$). In other cases, we may get multiple matching.

Types of alignments [2]. From the philosophical point of view, matching (alignments) of structural descriptions can arise from three specific sources. i) Due to prior knowledge, only some parts of the molecules are considered when establishing a matching. For example, backbone atoms of a protein must match backbone atoms of the other protein, when the 3-D structures are compared. ii) If we deal with unstructured descriptions, such as vectors, the elements – the vector components – match by definition, which is called canonical matching; iii) Finally, we might be interested in the maximal matching of two larger structures given in the form of structured descriptions (consisting of both entities and relationships), and then we need an optimizeable similarity measure such as in equation 5, in order to find a maximal alignment. The number of all possible alignments is very high, so finding the optimum is often very compute-intensive and sometimes intractable problem.

Algorithmic solutions From the algorithmic point of view, the methods can be subdivided a) according to the structure types (character strings, graphs, 3D structures), b) according to the nature of the matches that are being sought (exact matching, approximate matching), or according to number of partners compared (pairwise alignments or multiple alignments).

The majority of algorithms can cope only with the simplest descriptions, character sequences. Finding exact matches between two sequences of n and m characters respectively, has a complexity of $O(n,m)$. Comparison of graphs is much more difficult, here the majority of the problems are NP complete, i.e. computationally not tractable. Identity of graphs is determined by graph isomorphism algorithms, and similarity of graphs (such as protein structures, metabolic pathways) is a subgraph isomorphism problem, which is more difficult, and is aggravated by the fact that the structures in question are labelled graphs. Rigorous comparison of complex descriptions such as 3D structures is np-hard.

Time raqirements can be decreased if we use unstructured descriptions that do not need an alignment for comparison. These descriptions are simpler, and perform well only if one can find an adequate resolution, such as a multidimensional vector. The calculations are fast, but there is no guarantee that the results will be similar to those produced by alignment methods. On the other hand, such methods are useful as preliminary filters used to screen large databases.

Heuristic solutions provide the second general avenue for diminishing computer times. Most of the alignment methods use some kind of heuristics. There are two important heuristics that are used to simplify the process of alignment, the principle of linearity and the use of higher order descriptions.

The *principle of linearity* is based on the chain-like topology of protein and DNA molecules. All biologically important alignments contain short stretches that are very similar or identical between the two molecules. So instead of testing all possible

alignments, one can start identifying the highly similar chain segments and then combining them into larger alignment, which is computationally much less expensive. This philosophy it can be used both for sequences and for 3-D chain. Let *A* and *B* be polypeptide chains of *l* and *k* residues, respectively, and A_i^n denote a contiguous fragment of *n* residues of protein *A* starting at residue *i*. In this case, $[A^n] = A_1^n,\ A_2^n \ldots A_{l-n+1}^n$ will be an ordered list of overlapping fragment descriptions covering the entire chain of protein *A*. Let's provide such a list for both proteins and compare the fragments using a proximity measure $PM_{i,j} = PM(A_i^n, B_j^n)$. *PM* must be a proximity measure that can be unequivocally determined for any two fragments. In most cases this means that no alignment is needed between fragments compared (alignment and gaps would make the process prohibitively expensive). In some cases the precomputed or *a priori* known values of *PM* are stored in lookup tables. $PM_{i,j}$ values define a so-called similarity matrix, which is a symmetrical matrix of *k* x *l* elements (more accurately (l-n+1)x(k-n+1) elements). If $PM_{i,j}$ is a similarity measure, similar segments within two proteins appear as series of large values parallel to the diagonal of the matrix. The similarity matrix is used – under various names – for the determination of an overall alignment in several algorithms, many of which use dynamic programming techniques. Global alignments that extend from the beginning to the end of both sequences are found via an exhaustive search for the maximal matching, based on such methods as the Needleman-Wunsch algorithm [38]. Local alignments can be found via similar strategies, such as the Smith-Waterman and the Sellers [39] algorithms, as well as by heuristic solutions such as the FASTA and the BLAST algorithms.

All of these algorithms were developed for sequence alignment, where the fragments are overlapping *n*-words of amino acids, the scoring is be based on a sequence similarity score such as in equation 5. Naturally, one can also use 3-D description of the backbone for longer peptide segments of 3D structures, and use the *rmsd* distance for comparison. The actual algorithms are problem dependent, further examples are given in section 4.

The principle of higher order descriptions is based on the simple fact that comparing a smaller number of higher-order elements takes less time. The best example is the comparison of protein structures in terms of secondary structure elements. In addition to decreased computer time, higher order descriptions, such as secondary structure elements incorporate a great deal of human knowledge. As a consequence, the results of comparisons are usually close to human understanding.

Finally we mention that alignment is an optimisation problem, so all optimisation algorithms can be used for aligning structures. The optimum is understood in the context of the chosen representation and scoring scheme and may involve parameters that have to be adjusted on an empirical basis. Most users would therefore agree that alignments produced by computer programs can always be improved upon visual inspection.

4. Similarity spaces

In the foregoing, we reviewed the mathematical concepts relevant to the definition of similarity in bioinformatics: equivalence, matching, partial ordering, and proximity. These relationships arise in the context of a mathematical space. A mathematical space suitable for molecular similarity analysis is called a *molecular similarity space* and is defined to consist of a) a set of mathematical representations of molecules and b) one or more similarity relationships defined on this set. For example, one of the possible protein similarity spaces contains the sequences as representations, plus a set of equivalence

classes, each containing members of a protein family. It is assumed that a sequence similarity measure is also defined on the set of sequences. Another similarity space used for proteins consists of the structures or protein folds as descriptors, a set of equivalence classes, each containing members of a specific fold group. A distance function, such as *rmsd* is defined on the set of fold structures.

The co-existence of *a priori* known (biologically relevant) classification schemes and computable proximity measures is characteristic of the similarity spaces studied by bioinformatics. In the typical case, the database also contains a large number of unclassified objects (sequences, structures), and much effort is put into either founding new classes for some of these objects, or trying to fit them into one of the existing categories. It is noted that a proximity measure can be used to establish a computable classification using one of the many clustering methods. In a fortunate case the computed clustering is consistent to the *a priori* known classification, and the potential new clusters that have no *a priori* known counterparts are excellent candidates for discovering new, biologically relevant classes.

Methods for representing *a priori* known categories can be grouped according to the nature of description used for the individual categories [40]. Classical summary descriptions are *consensus descriptions* that are valid for all members of a category. Probabilistic summary descriptions are valid only with some probability. Consensus descriptions such as sequence patterns can be pictured as the description of a prototype in the given class. In contrast to consensus descriptions, *exemplar-based descriptions* represent the categories as a database consisting of the members of the category. All of these methods have been used e.g. for protein domain sequences. Domain sequence collections and domain annotations in protein sequence databases are exemplar-based descriptions. Regular expressions are classical summary (consensus) descriptions that are supposed to be valid for all members, and there is a variety of statistical (probabilistic) descriptions [40].

The problem of classification is one of the fundamental exercises in such fields a domain sequence identification, or function prediction. Given a set of classes A^i in a database, the classification of a sequence is often based on minimal distance (or maximum similarity). Oftentimes, the class A^i of the closest object $[\min_{i,j} PM(S, A_j^i)]$ is automatically assigned to an unclassified object. In other cases, the closest class is determined from the consensus-representations of the classes, using $\min_i PM(S, < A >^i)$. The use of mathematical spaces in the analysis of chemical structures is reviewed in [2, 18].

5. Conclusions

Summarizing we can conclude that the description of structures as entity-relationship networks provides a simple framework to describe the use of similarity in various fields. There are a number of qualitative concepts, such as similarity groups (equivalence classes), patterns as well as quantitative concepts, such as similarity measures that are present in all fields. Mathematical spaces ("similarity spaces") provide a way for describing databases as well as the mathematical tools of analysis in a common framework. The definitions listed in this review are applicable in other fields of bioinformatics not explicitly mentioned in this review, such as the analysis semantic similarities [9] or the analysis of networks [41]. An overview of practical applications will be published in a subsequent chapter in this volume [8].

The description of structures as entity-relationship networks provides a simple framework to describe the use of similarity in various fields. There are a number of

qualitative concepts, such as similarity groups (equivalence classes), patterns and quantitative concepts, such as similarity measures that are present in all fields.

Acknowledgements

This material 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 to the following individuals for their comments on various topics in the manuscript: Stephen Altschul (Bethesda, US), Steve Bryant (Bethesda, US), Alexandre De Leon, (Calgary, Canada), Jacques Demongeot (Grenoble, France), Mark Gerstein (New Haven, UK), Andrew Harrison (London, UK), Lisa Holm (Hinxton, UK), Jack Leunissen (Wageningen, The Netherlands), Christine Orengo (London, UK), William F. Pearson (US), János Podani (Budapest, Hungary).

References

[1] Pongor, S., *Novel databases for molecular biology.* Nature, 1988. **332**(6159): p. 24.
[2] Johnson, M.A. and G.M. Maggiora, *Concepts and applications of molecular similarity*. 1990, New York: Wiley-Interscience. 393.
[3] Kanehisa, M., *Post-genome informatics*. 2000, Oxford New York: Oxford University Press. 148.
[4] Baldi, P. and S. Brunak, *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. 2001, Cambridge, MA: MIT Press. 400.
[5] Durbin, R., et al., *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. 1999, Cambridge: Cambridge University Press. 368.
[6] Ripley, B.D., *Pattern Recognition and Neural Networks*. 1999, Cambridge: Cambridge Univeristy Press. 403.
[7] Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. 2000, Cambridge: Cambridge University Press. 189.
[8] Vlahovicek, K., et al., *Concepts of similarity in bioinformatics: Principles of applications to sequences, protein 3D structures and genomes.*, in *Introduction to Bioinformatics*, S. Jelaska and D.S. Moss, Editors. 2003, Kluwer Academic Publishers: Boston, Dordrecht, London. p. in press.
[9] Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
[10] Carugo, O. and S. Pongor, *The evolution of structural databases.* Trends Biotechnol, 2002. **20**(12): p. 498-501.
[11] Csányi, V., *Evolutionary Systems and Society*. First ed. Vol. 1. 1989, Durham and London: Duke University Press. 257.
[12] Kampis, G., *Self-modifying systems in Biology and Cognitive Science*. First ed. International Series in Systems Science and Engineering, ed. G.J. Klir. Vol. 1. 1991, Oxford, New York: Pergamon Press. 543.
[13] Hátsági, Z., V. Skerl, and S. Pongor, *Motifs in Protein Sequences: Towards a unified view on sequence databases*, in *Biotechnology Computing*, L. Hunter, Editor. 1994, IEEE Computer Society Press: Los Alamos, CA. p. 255-264.
[14] Ashburner, M. and S. Lewis, *On ontologies for biologists: the Gene Ontology--untangling the web.* Novartis Found Symp, 2002. **247**: p. 66-80; discussion 80-3, 84-90, 244-52.
[15] Goldmeier, E., *Über die Ähnlichkeit bei gesehenen Figuren.* Psychol. Forsch., 1936. **21**: p. 146-208.
[16] Goldmeier, E., *Similarity in visually percieved forms*. 1 ed. Psychological Issues, ed. H.J. Schlesinger. Vol. 29. 1972, New Yorik, N.Y.: International Universities Press, Inc. 135.
[17] Gentner, D., *The mechanisms of analogical learning*, in *Similarity and Analogical REasoning*, S. Vosniadou and A. Ortony, Editors. 1989, Cambridge, University Press: Cambridge, U.K. p. 199-241.
[18] Johnson, M.A., *A review and examination of mathematical spaces underlying molecular similarity analysis.* Journal of Mathematical Chemistry, 1989. **3**: p. 117-145.

[19] Sali, A. and T.L. Blundell, *Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming.* J Mol Biol, 1990. **212**(2): p. 403-28.
[20] Sali, A., et al., *From comparisons of protein sequences and structures to protein modelling and design.* Trends Biochem Sci, 1990. **15**(6): p. 235-40.
[21] Via, A., et al., *Protein surface similarities: a survey of methods to describe and compare protein surfaces.* Cell Mol Life Sci, 2000. **57**(13-14): p. 1970-7.
[22] Via, A., et al., *Three-dimensional view of the surface motif associated with the P-loop structure: cis and trans cases of convergent evolution.* J Mol Biol, 2000. **303**(4): p. 455-65.
[23] Pawlowski, K. and A. Godzik, *Surface map comparison: studying function diversity of homologous proteins.* J Mol Biol, 2001. **309**(3): p. 793-806.
[24] Ankerst, M., et al., *Nearest neighbor classification in 3D protein databases.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 34-43.
[25] Sneath, P.H. and R.R. Sokal, *Numerical Taxonomy*. 1973, San Fransisco: Freeman. 256.
[26] Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallogr. A, 1976. **32**: p. 922 –923.
[27] Carugo, O. and S. Pongor, *A normalized root-mean-square distance for comparing protein three-dimensional structures.* Protein Sci, 2001. **10**(7): p. 1470-3.
[28] Johnson, M.S. and J.V. Lehtonen, *Comparison of protein three-dimensional structure*, in *Bioinformatics. Sequence, structure and databanks*, D. Higgins and Taylor, W., Editors. 2000, Oxford University Press: Oxford New York. p. 15-50.
[29] Rossmann, M.G. and P. Argos, *Exploring structural homology of proteins.* J Mol Biol, 1976. **105**(1): p. 75-95.
[30] Irving, J.A., J.C. Whisstock, and A.M. Lesk, *Protein structural alignments and functional genomics.* Proteins, 2001. **42**(3): p. 378-82.
[31] Willett, P., *Similarity and clustering in chemical information systems*. 1987, New York: John Wiley & Sons Inc. 254.
[32] Bajic, V.B., *Comparing the success of different prediction software in sequence analysis: a review.* Brief Bioinform, 2000. **1**(3): p. 214-28.
[33] Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-24.
[34] Murvai, J., K. Vlahovicek, and S. Pongor, *A simple probabilistic scoring method for protein domain identification.* Bioinformatics, 2000. **16**(12): p. 1155-6.
[35] Murvai, J., et al., *Prediction of protein functional domains from sequences using artificial neural networks.* Genome Res, 2001. **11**(8): p. 1410-7.
[36] Carugo, O. and S. Pongor, *Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison.* J Mol Biol, 2002. **315**(4): p. 887-98.
[37] Evans, M., N. Hastings, and B. Peacock, *Statistical Distributions*. 3nd edition (June 15, 2000) ed. 2000: John Wiley & Sons. 221.
[38] Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
[39] Sellers, P.H., *The theory and computation of evolutionary distances.* Journal of Algorithms, 1980. **1**: p. 359-373.
[40] Smith, E.E. and D.L. Medin, *Catgories and Concepts*. Cognitive Science Series. 1981, Cambridge, MA: Harvard University Press. 203.
[41] Dorogovtsev, S.N. and J.F.F. Mendes, *Evolution of Networks*. 2003, Oxford: Oxford University Press. 264.