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

Biology and Informatics

Alan L. MACKAY

School of Crystallography, Birkbeck College, University of London[1], Malet Street, London WC1E 7HX
(The Inter-University Centre, Dubrovnik, 19-24 May 2003)

Abstract. The advent of modern bioinformatics is the result of a long succession of scientific discoveries and paradigm changes in chemistry and biology. This chapter provides an introduction to the pertinent events in these diverse fields.

Introduction

"Is it not a wonder that anyone can bring himself to believe that a number of solid and separate particles by their chance collisions and moved only by the force of their own weight could bring into being so marvellous and beautiful a world?" Marcus Tullius Cicero[2] (106-43 BC), *"On the Nature of the Gods"*

"Molecular Biology is the confluence of information and conformation" John Kendrew, (1965)

"How does so little information control so much behaviour?" Richard L. Gregory, in *Towards a Theoretical Biology*, (ed. C. H. Waddington), (1969)

"In less than a generation we have witnessed a radical, irreversible, world-wide transformation in the way that science is organised, managed and performed." John Ziman, *"Real science: what it is, and what it means",* CUP 2000, (p.67).

1. Atomism

The city of Dubrovnik, earlier The Republic of Ragusa, in which this workshop was held, is a historic place and we have to mention its most famous scientist, Roger Joseph Boscovich[3] (1711-1787, FRS (1761)), who was born here and who looked after its interests, although he usually resided elsewhere. He worked mostly on astronomy, but he was an atomist and had proposed an important theory of point atoms, between which were mutual forces with a number of minima at different distances, running from a strong repulsion at very short distances, to an inverse square attraction like gravitation at very long distances[4]. This removed difficulties about what happens at the discontinuity of the surface of a billiard-ball kind of atom. Boscovich, who was a Jesuit, lived a generation after Newton and influenced Maxwell and Kelvin among others. He aimed to understand the properties of things in terms of their structure and his main work was called *"Philosophiae naturalis theoria reducta ad unicum legem virium in natura existentium",* (Vienna, 1758, etc.) [Physics reduced to a single law of the forces existing in nature]. Also in Dubrovnik, Marin Getaldic[5] (1568-1626), a century earlier, appears as a pioneer of the algebraic geometry which is the basis of computer graphics. Already two and three hundred years ago European scientists were remarkably closely in touch with each other.

A generation before Boscovich, Newton, having determined "the motions of the planets, the comets, the Moon and the sea", was unfortunately unable to determine the remaining structure of the world from the same propositions because, as Newton said:

"I suspect that they may all depend upon certain forces by which the particles of the bodies, by some causes hitherto unknown, are either mutually impelled towards one another, and cohere in regular figures, or are repelled and recede from one another. These forces being unknown, philosophers have hitherto attempted the search of Nature in vain; but I hope the principles laid down will afford some light either to this or some truer method of philosophy". *(Preface to the Principia).*

But neither Newton nor Boscovich had the present-day experimental access to the atomic and molecular level necessary for the understanding of chemistry and biochemistry.

Boscovich was just one in the long tradition of atomism, which had started with Leucippus and Democritos, and which was promoted by Lucretius in his Latin poem, "On the Nature of the Universe" which sought to explain everything in terms of atoms[6]. Lucretius specifically claimed that mind and spirit are both also made of atoms. Atomism has long been a difficulty for the Vatican, most recently in connection with allergy to gluten and its implications for the doctrine of transubstantiation, but Lucretius' programme is steadily becoming reality.

Atoms became visible after the proof of their arrangement in crystals of sodium chloride by Lawrence Bragg and William Bragg, following the discovery of X-ray diffraction in 1912 by Laue, Friedrich and Knipping. Explaining everything in terms of atoms then became a major feature of modern science, especially of molecular biology[7].

2. Towards a theoretical biology

In the 1920s and 1930s around the laboratory of F. Gowland Hopkins ("the father of biochemistry") in Cambridge, there flourished the Club for Theoretical Biology, which was a most important source of ideas about molecular biology. The key idea was that the three-dimensional structure of molecules determined their behaviour. The group included Joseph Needham, Conrad Waddington, Desmond Bernal, Lancelot Whyte[8], and others. They aimed to make biology a real science like physics, where there were interactions to and fro between theory and experiment, and to understand the origin and processes of life. They also had radical political ideas. Theoretical biology was a new concept. Darwin had formulated the principles of evolution by natural selection, but now there was a prospect of elucidating the mechanisms of heredity, which appeared to operate at the atomic level.

Needham made a proposal (1935) for an Institute of Physico-chemical Morphology (to the Rockefeller Foundation, through Warren Weaver) but this was not funded, although Weaver and Astbury (independently) had coined the expession 'Molecular biology'. Needham as an embryologist had to ask how shape and the unfolding of shape in the embryo following a programme, was determined by the hereditary material.

Already at that time (1931) Bernal[9] had recognised that in order to be replicated, the hereditary material, then thought to be protein, had to be a linear structure. It was not demonstrated until 1944 that genes were nucleic acids and not protein (although associated with protein)[10]. In 1934 Bernal showed that if a crystal of pepsin were kept in its mother liquor, the diffraction pattern had information out to inter-atomic dimensions, that is, a protein molecule has every atom in its proper place. This was an epoch-making discovery. The concept of a mystical protoplasm thus collapsed. Proteins had a structure which could be investigated by physical methods, chiefly X-ray crystal structure analysis. Bernal and Astbury agreed to divide the new world between themselves, Bernal taking the globular proteins and Astbury the fibrous. There are several excellent studies of the development of molecular biology[11]. Schroedinger's well-known book "What is life" (1945)[12] appeared rather late in the day.

C. H. Waddington (1905-1975), a biologist, one of the Cambridge Club, after the war organised a series of influential seminars under the title of "Towards a Theoretical Biology" (1968, 1969, 1971) which brought together varied people who digested the revolution in computing, information, the structure of molecules, genetics and the origin of life. He himself promoted the concept of the "epigenetic landscape" as a way of visualising the development of an organism as genes were switched on and off to make choices between various paths. Almost all the people concerned with protein structure had exceptionally well-developed abilities for spatial visualisation (now shown by PET scanning to correspond to physical development of structures in the brain. Information and thought really have a material basis as Lucretius had suspected).

3. Hierarchy

The great success of X-ray crystal structure analysis in providing the shapes of molecules, has obscured the fact that most materials are not crystalline, although almost everything gives useful X-ray diffraction patterns. Crystallisation is a test for purity, but crystals are exceptional in that one rule takes one from the atomic level of 1 Angstrom (0.1nm) right up to 10 cm. The span of operation of this rule is unusually great. The recent discovery of quasi-crystals has led to a profound re-assessment, leading in the direction of hierarchy, of the laws of crystallography.

Biological structures are distinctively hierarchic with perhaps six levels of organisation with much smaller spans, each with its characteristic rules of ordering. These levels overlap to a greater or lesser extent. Properties at one scale are determined by structure at that scale, but may be critically influenced by certain detailed configurations in the level below for which the level above forms an average climate.

Levels of organisation or integration were clearly recognised by, for example, Joseph Needham[13], representing the thought of the Club for Theoretical Biology.

4. Information theory and the computer. Information and material structure

The concept of information began to appear in the 1920s. Not surprisingly, information theory began with the question: "How much should you pay for your telegraph message and how fast would it go?" At first it was so much a word but then newspaper correspondents began to make up pseudo-words like "Pariswise urgentmost". Theory began to be developed for questions of military cryptography, as the story of the Enigma machine has revealed. The Colossus computer[14] was built for cryptography at Bletchley Park. Questions of bandwidth arose. How much information could be transmitted over a land-line? The first Atlantic cable could only carry a few bits per second. Nyquist (1924), Kolmogorov and Hartley (1928), Claude Shannon, Louis Brillouin, Warren Weaver, Leo Szilard, Norbert Wiener were all concerned with the foundation of information theory[15].

John Tukey invented the word "bit" for a binary digit and Shannon used the word "entropy" for information content. as $-\Sigma\, p_i \log p_{i.}$ (where p_i is the fractional probability of the i-th kind of character. There is still great confusion as to the entropy content of meaningless and meaningful information. Shannon's example was of printed English text and he showed that about half the information is arbitrary, that is, is "meaning", and half is redundancy due to the intrinsic structure of the language which every native speaker knows. This redundancy can be used to correct mistakes in transmission. The Huffmann algorithm for compression[16] is based on a knowledge of the relative probabilities of different symbols, measured over a particular text. Since Shannon, the analogy between DNA and protein sequences and natural languages has been pervasive.

Information theory was developed in dialogue with the construction and use of computers which have made both the examination of the arrangement of atoms and the operation of data-bases possible. “Cyberspace” was invented and colonised the literary world[17].

Donald Booth at Birkbeck, recruited by Bernal to make a computer for crystallography, invented the floppy disc[18], using a primitive speech recorder with a magnetic disc, but he discarded it, and toyed with the machine translation of natural languages, an idea, which emerged in discussions with Warren Weaver. The Cambridge Crystal Structure Database was begun in an attic at Birkbeck College, originally on cards, before being established in Cambridge. Its creation was due to Olga Kennard and J. D. Bernal (who had far earlier been concerned with the development of Structure Reports (originally Strukturbericht) collecting all data on the arrangement of atoms in crystals.

Gregory Chaitin proposed that the amount of information in a structure could be defined in terms of the shortest computer programme necessary to generate it. The number of operations necessary to sort a sequence of N numbers into an arbitrary order is N log N ("operation" needs more careful definition).

5. Cellular automata

Robert May alerted us to the fact that there were many “simple mathematical models with very complicated dynamics”, although the immensely creative J. B. S. Haldane had noted this more graphically in 1932[19]. In particular, finite difference equations, for example $x_{t+1} = f[x_t]$, have results which cannot be predicted far ahead better than by simply iterating the process. It also emerges that eventually the finite accuracy of all computing processes, including those in nature, will render the outcome indefinite and unpredictable. This kind of equation can be extended to two or three (or more) dimensions, the equations may be coupled or non-linear, so that the complexity increases. Stephen Wolfram[20] has developed certain classes of “cellular automata” in such detail that classification is possible. Intriguing and unpredictable patterns may emerge[21]. It is immediately clear that patterns in nature, particularly those in biological systems produced by the switching on and off of genes which synthesise proteins, must be physically analogous to such mathematical phenomena. Now even the classical mechanical problems of Newton, the pendulum and the solar system are seen to be weakly chaotic.

6. Structural molecular biology. Proteins and nucleic acids

Desmond Bernal had the good fortune to be the right man in the right place at the right time. In February 1945, before returning to Birkbeck after the war, Bernal produced a plan *"to set up a research centre for the study of the structure and properties of large molecules by all available physical and chemical methods".* This was based directly on the thinking of the Cambridge club and was effectively the charter for the Birkbeck Laboratory, set up in 21-22 Torrington Square, which Bernal headed from 1938 to about 1964. In the 1950s

Aaron Klug, Rosalind Franklin, Kenneth Holmes and others contributed greatly to the establishment of molecular biology. I do not need to list their enormous achievements.

If we take a large molecule, for example the protein lysozyme, it contains C, H, O, N, S atoms in definite numbers and so should appear as a region in the phase diagram of this 5-component system. It would probably be in a meta-stable energy minimum. However, this is clearly unrealistic and lysozyme is much better considered as being specified by a number which represents its amino-acid sequence, which is effectively its address in phase space. Given the sequence, lysozyme can be now be made by adding the right amino residues in the right order. That is, it has a description. Information is stored in such meta-stable systems.

The proteins of life are a very special and minutely small subset of all possible amino-acid sequences characterised by being able to fold up into a unique configuration.

7. The double helix

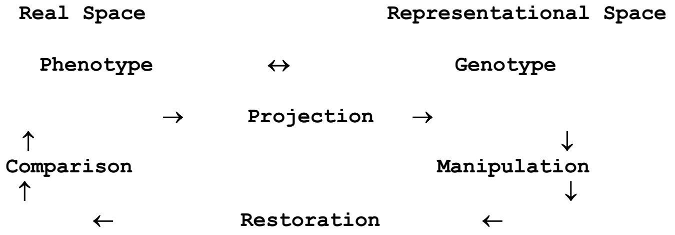
Darwin and Mendel recognised the discrete nature of the hereditary substance but could get no further without access to the levels below those provided by optical microscopy.

The 50th anniversary[22] of the spatial structure of the DNA double helix has ensured that the circumstances of the discovery should now be well-known. I remember opening the copy of Nature for 25 April 1953 and reading the three papers disclosing the double helix, Crick and Watson's paper ending with the sentence: *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"* and thinking, yes, of course, it must be something like that. An immediate entry to the mechanism of heredity had opened, just as van't Hoff's vision in 1874 of the tetrahedral carbon atom had opened up organic chemistry, making it clear that is was the arrangement of atoms in three-dimensional space that was the determining factor for molecules (although Pasteur had demonstrated optical enantiomorphism in 1848, picking left- and right-handed crystals, and this implied spatial structure).

8. Dialectics

The key theme of this workshop is the relationship between information and structure. The more we look into it, the more complicated it gets. A very informative modern survey of the interaction between *"nature and nurture"* has been provided by Matt Ridley[26]. The conflict has been fought at all levels from the molecular to the politics of agriculture and education. The Lysenko affair in the Soviet Union was one acute manifestation, but there are still deepening conflicts with religious views.

The basic idea, that one structure should be a description of another, but both composed of atoms subject to the same laws of chemistry, has been revolutionary[23].



Earlier philosophical systems analysed economics and society as equilibrium systems, in many cases fixed by the unchanging dogmas of sacred texts. Change can now also be handled explicitly. Newton and Leibnitz, with the differential calculus, provided the tools for physics and Hegel introduced the idea of dialectics into philosophy[24]. In science there are many new ways of handling change, for example, the epigenetic landscape of C. H. Waddington for biology, all kinds of computer simulations of systems ranging from the Solar system (found to be weakly chaotic) to the British economy. Arthur Winfree, a pioneer in dealing with non-linear systems, gave his book the intriguing title “The geometry of biological time” indicating that changes in time and space were intimately mixed (as also in relativity) although Joseph Needham had much earlier written: "form is simply a short time-slice of a single spatio-temporal entity". The sudden changes in such systems have been illustrated in the ‘catastrophe theory’ of René Thom which has been expounded by Christopher Zeeman for social as well as for physical systems. They find that there are only seven types of geometrical singularities in the configuration space. These are by way of being mathematisations of the “double bind” kind of situation which philosophers describe where one can get out of a knot only by jumping to some other position.

9. Experimental techniques

Of course the whole progress of bioinformatics has depended on the development of experimental methods and their implementation, both facilitated by the advent of computer hardware and of appropriate algorithms. Structural studies stand on X-ray crystal structure analysis, electron microscopy, atomic force microscopy and nuclear magnetic resonance and all their variants.

Fred Sanger in Cambridge quietly provided the methods for sequencing both proteins (1949-55) and DNA, which are the absolute fundamentals for bioinformatics, but the flood of sequence data is a result of the industrial-scale implementation of sequencing methods on a huge scale[33].

Numerous automated biochemical techniques for identification and, for example, for combinatorial chemistry, have become essential.

Computer handling of gigantic data-banks and computer modelling of the conformations of proteins and the expected chemical properties of molecules are now central to bioinformatics.

10. Genomics

The key problems[25] include:

- The structure of proteins, protein folding, the operation of proteins.
- DNA, its sequence, replication, transcription, its interaction with proteins, the switching of genes.
- The ribosome.
- The structure and operation of chromosomes, meosis, mitosis, replication, mutation, variation. Extra-chromosomal nuclei acid.
- Phylogenetics, evolution, speciation.

The nematode worm, *Caenorhabditis elegans,* with 302 nerve cells*,* was the essential link, chosen by Sydney Brenner, between behaviour, molecules and genetics.

11. Genetic and financial engineering

The nature/nurture interaction means that the results of the expression of genes as proteins depend on the environment in which they are expressed[26]. Analogously, the consequences of the developments of biotechnology depend on the social system within which they are expressed. There are huge possibilities of good or evil. Thus, scientists cannot be unconcerned with politics and must act responsibly. We may be sure that some people, somewhere[27], are thinking about the uses of bioinformatics for military, selfish and destructive ends[28]. Social control of applications of genomics cannot be left to oligarchies[29]. This means that scientists in genomics must work to create an informed public and this implies an opposition to secrecy[30].

Today, information, especially that relating to molecular structure and genetic sequences, is being enclosed, as land was enclosed in the 18th and 19th centuries in Britain, and is becoming private property[31] (as are also computer components, algorithms and methods[32]). The Human Genome Project has generated acute conflicts in the "Republic of Science" and more generally[33]. Huge data-banks of the DNA information on individuals are being built up for social, political and military purposes. Even the ownership of ordinary standard English words and phrases are being claimed by arrogant companies and litigation[34] absorbs a large proportion of the social product, especially in the USA.

With the development of socio-biology, through the efforts of E. O. Wilson, D. S. Wilson, R. Dawkins, J. Goodall and many others, the extension of biological ideas, from the collective behaviour in insect societies to the noösphere, is making progress towards understanding the behaviour and evolution of individuals, groups and species. “Memes” have been proposed as units of social structure[35] circulating in the world of information. At the insect level some quantitative confirmation of numerical predictions has been achieved. Such topics should eventually be included in bioinformatics as part of the dialogue between information and matter.

12. Lucretius

In due course I and you and everyone else will cease to operate as living systems. The atoms will disperse and all that will be left will be traces of information distributed round the world. There will be bits of genetic sequence continuing in descendants and in relatives, some genetic information may be recoverable from organic specimens, there will be items in the Internet and in documents of all kinds. There will also be transient memories residing in others. It will all be a matter of chance as to what survives of us, but it will be information recorded in various kinds of matter.

I must draw your attention again to Lucretius' book: *De Rerum Natura*, [on the nature of things]. It reached us from antiquity in only a single manuscript copy and, with the development of printing at the end of the fifteenth century, it was reprinted and translated and generally circulated so that, by chance, this remarkable philosophic outlook has survived to our own times[36] and remains a source of inspiration and consolation for us even two thousand years later.

It is information transmitted in code from our ancestors and it is this coding into language which distinguishes the human species from all others. Lucretius had said "I set out to loose the mind from the knots of religion"[37]. His book’s great merit is that it sought to give a complete, unitary picture of the universe, free from prevailing superstitions. Also, we might note, prophetically perhaps, in view of the concern over AIDS and SARS (Severe

Acute Respiratory Syndrome) that Lucretius ended his book with a description of the social chaos which occurred with the plague in Athens. I commend it to you as the foundation of bioinformatics.

13. The Present Crisis

This workshop takes place at a critical time in human history[38]. Science and technology have changed the world[39]. We cannot avoid the political significance of bioinformatics and indeed the militarisation of science must be one of our major concerns[40].

The human race faces the possibility of various catastrophes, from oligarchies to chaos, as well as natural disasters, most of its own making. In particular, the growth of the world population cannot continue indefinitely at its present rate. The only way in which these can be avoided is by knowledge and the intelligent application of knowledge[41]. Thus it is vital to build a world-wide network of people who understand each other, who have each other's confidence, who can operate in their own societies, and who will be able to inject their special knowledge into the decision-making centres[42] and thus to influence the course of history. The social parts of our meetings, begun in Dubrovnik, are at least as important as the technical parts.

Notes

[1] [London] … "the quick forge and working-house of thought" *W. Shakespeare, King Henry V, 5:23.*

[2] Cicero was supposed to have been the editor of Lucretius' m/s "De rerum natura".

[3] L. L. Whyte (ed.) *"Roger Joseph Boscovich",* Allen and Unwin, London, 1961.

[4] G. Malescio, "Intermolecular potentials – past, present and future", Nature materials, **2**, 501-503, (2003).

[5] "De resolutione et compositione mathematica", Rome (1630). Getaldic also made a concave parabolic mirror, 70 cm in diameter, which is now in the London Maritime Museum (inventory NAV 0928), and probably also a reflecting telescope.

[6] Lucretius, [Titus Lucretius Carus], "On the Nature of the Universe", (trans. R. E. Latham), Penguin, revised edition 1994. In this book, the concept of "swerve", [clinamen] which has so worried the classical commentators, can perhaps be understood retrospectively in terms of chaos theory where the progress of an idealised game of billiards cannot be forecast more than a few impacts ahead.

[7] A. L. Mackay, "Generalised crystallography", Structural Chemistry, **13,** (3/4), 217-222, (August 2002). http://sinapse.arc2.ucla.edu/Mackay02.pdf

[8] The editor of the work on Boscovich and a writer on atomism.

[9] "The facts of genetics demand, as J.B.S. Haldane has pointed out, that, at some stage in mitosis, the individual molecules in a chromosome must be exactly duplicated. A complete molecule can be duplicated in three ways. If it is solid and three dimensional only a supernatural agency, a divine copyist, can, entering its inner complexity, reproduce it in detail. If we prefer a natural solution, we must imagine the molecule stretched out either in a plane or along a line. In either case the simpler constituent molecules have only to arrange themselves one by one on their identical partners in the original molecule, and then become linked to each other by the absorption of suitable quanta from radiation or from second order collisions. That such autocatalysis is possible is indicated by recent work in Russia and America, where the regular atomic arrays of metallic catalysts are shown to operate like laceworker's frames on which simple organic molecules settle to be joined into larger aggregates. A two-dimensional reproduction of this kind is impossible, owing to the fact that the constituent amino acids in nature are not symmetrical, but exist in right or left hand forms. Two-dimensional reproduction would lead to mirror image molecules, which are not found in nature. There remains then only one dimensional reproduction. At the moment of reproduction, but not necessarily at any other time, the molecule of the protein must be imagined as a pseudo-linear, associating itself, element by element, with identical groups, related by an axis instead of a plane of symmetry, and thus preserving only right – or only left handed symmetry. This hypothesis is clearly indicated by Astbury's explanation of Svedburg's numbers. Svedburg has established that most natural proteins consist of M Wt 34,000 or multiples 2, 3, or 6 times that number. This gives us the confidence to treat all protein molecules, regardless of their

complex constitution, as belonging to one natural species. It is impossible to claim that these ideas are anything but preliminary guesses, but they have the advantage of being susceptible to experimental test."
J. D. Bernal (1931) [Int. Congress of the History of Science. Bernal Archive, Cambridge. A4.7 Box 22, by courtesy of Andrew Brown].

[10] Philip Ball, "Portrait of a molecule", Nature, 421, 421-422, (23 January 2003).

[11] H. F. Judson, *"The Eighth Day of Creation",* Simon and Schuster, New York, 1979.
Robert Olby, "The Path to the Double Helix ", Macmillan, London,
Nature, **421** , (6921), (23 Jan. 2003) [special supplement for the 50th anniversary of the double helix]

[12] Schroedinger used the term "aperiodic crystal" which later entered the discussion of quasi-crystals after 1985. He said: "We believe a gene – or perhaps the whole of the chromosome fibre – to be an aperiodic solid".

[13] J. Needham, *"Order and Life",* (1936) Reprinted MIT Press, 1968. [Dedicated to the Theoretical Biology Club.]

[14] The Colossus computer, all copies of which were destroyed after the war on Churchill's orders, is now being rebuilt at Bletchley Park as an historic monument.

[15] L. Brillouin, "Science and Information Theory", New York, 1956.
C. E. Shannon and Warren Weaver, *"The Mathematical Theory of Communication",* University of Illinois Press, (1949).
D. M. Mackay, "Quantal Aspects of Scientific Information", Phil. Mag., 41, (1950) and Proc. First London Symposium on Information Theory, (1950)

[16] A. L. Mackay, "Optimisation of the genetic code", Nature, **216**, 159-160, (1967).

[17] "Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphical representation of data abstracted from the banks of every computer in the human cystem. Unthinkable complexity. Lines of light ranged in the non-space of the mind, clusters and constellations of data. Like city lights, receeding... " William Gibson (ca. 1982).

[18] A. D. Booth, "A magnetic digital storage system", Electronic Engineering (July, 1949)

[19] J. B. S. Haldane (1892-1964) wrote, very presciently,
"Even in a non-mathematician like myself, some differential equations evoke fairly violent physical sensations similar to those described by Sappho and Catullus when viewing their mistresses. Personally, however, I obtain an even greater 'kick' from finite difference equations, which are perhaps more like those which an up-to-date materialist would use to describe human behaviour". Haldane was, indeed "an up-to-date materialist"! *"The Inequality of Man*", (1932), Penguin, (1937), p. 39.
Robert M. May, Nature, **261**, 459-, (10 June 1976).
(http://nedwww,ipac.caltech.edu/level5/Sept01/May/May_contents.html)
See also: A. L. Mackay, Physics Bulletin, 495-497, (Nov. 1976) and Izv. Jugoslav. Centra za Krist., **10**, 25-36, (1975). (http://www.cryst.bbk.ac.uk/surfaces/zagreb.html). J. W. Galloway, Physics Bulletin, **34**, 161-164, (1983).

[20] Stephen Wolfram, *"A New Kind of Science",* Wolfram Media, 2002.

[21] For example, P. Ball, *"The self-made tapestry: Pattern formation in Nature"*, Oxford, (1999).

[22] Nature, **421**, (23 January 2003).

[23] A. L. Mackay, "From 'The Dialectics of Nature' to the inorganic gene", Foundations of Chemistry, **1**, (1), 43-56, (1999).

[24] I have discussed this at greater length in "From the 'Dialectics of Nature' to the inorganic gene", Foundations of Chemistry, **1**, 43-56, (1999).

[25] A. M. Lesk, *"Introduction to Bioinformatics",* Oxford University Press, 2002.
(www.oup.com/uk/lesk/bioinf/)

[26] Matt Ridley, *"Genome: the autobiography of a species in 23 chapters*", Fourth Estate, London, 1999.
"Nature via Nurture", Fourth Estate, London, 2003.

[27] Military uses of bioinformatics are discussed in: Tom Mangold and Jeff Goldberg, "Plague Wars", Macmillan, London, 1999.

[28] Concerns about anthrax illustrate this. The USSR had a serious accident releasing anthrax; the USA also had a dramatic terrorist attack associated with its own weapons programme; much earlier Churchill wished to use anthrax, tested on Gruinard Island in the North of Scotland, against the German civil population; Iraq too, had sought to develop anthrax.

[29] In Britain already two million DNA profiles are held in police records.

[30] If you are a scientist at an American research university like mine, you know what to do if you think you've hit on some technique or bit of knowledge that might have commercial potential. You go online to the university's technology transfer office, download an invention and technology disclosure form, and fill in the details. You have to do that because all such intellectual property (IP) discovered by this university's employees belongs to the university. If the local bureaucrats think there's something in it, they will file a

provisional patent and. after formally offering it to any government agency that funded the research – which usually declines – they will start hawking the IP about to see if any entrepreneurs or companies want to license it. Priority in your IP is protected at this stage, and you can now go ahead and publish if you wish, but eventually you may proceed to full (or utility) patent, where property rights are wrapped up more securely, and, while IP lawyers make fortunes from litigation about who in fact owns the property, basically the matter is now in the domain of formal law. If the university does manage to license the IP, you will get perhaps 35 per cent of the royalty stream. Or, if that's not enough for you, you can cut yourself free from academia and take your chances with the venture capitalists as an independent entrepreneur. - Steven Shaplin, (University of California at San Diego), London Review of Books, 6 March 2003, p.14.

[31]"Monsanto aim to control the world food supply", [London, Channel 4 TV. "DNA the story of life", 19:00, 15 March 2003] see also for example the website www.cryptome.com for the current applications of surveillance technology.

[32]L. Cranswick, "The potential power of 'software patents' to destroy crystallographic software", Crystallography News, (84), (March 2003). http://www.ccp14.ac.uk/maths/software-patents/

[33]J. Sulston and Georgina Ferry, “The Common Thread: A story of science, politics, ethics and the human genome” Bantam, (2002).
see the review of this by Robin McKie, The Observer 3 Feb. 2002. at www.guardianunlimited.co.uk/ (search for McKie).
Apparently J. D. Watson told Sulston “Venter wanted to own the whole genome the way Hitler wanted to own the world”.

[34]There are some 900,000 lawyers in the USA. In Japan, with a different social structure there are only 18,000. Science is now done with lawyers looking over your shoulder.

[35]Richard Semon (Munich) had proposed “mnemes to be the preserving principle in the interaction of organic events” and this idea was promoted by Ernst Haeckel.

[36]Karl Marx, as a young man, wrote his doctoral thesis (presented *in absentia* at the University of Jena) on a comparison of the philosophies of Democritos and Epicurus.

[37]"religionum animum nodis exsolvere pergo"; I. 932.

[38]R. Brenner, "Towards the precipice: the crisis in the US economy", London Review of Books, 25, (3), (6 Feb. 2003); Chalmers Johnson, "Who's in Charge" (Review of Daniel Ellsberg, "Secrets: A Memoir of Vietnam and the Pentagon Papers), (LRB same number: see the London Review of Books website www.lrb.co.uk). E. Hobsbawm, “Age of Extremes: The short twentieth century 1914-1991”, London, (1994).

[39]See, for example, Chapter III of "The Theory and Practice of Oligarchical Collectivism" by Emmanuel Goldstein, (1949).

[40]War also is being privatised as Eisenhower's 'military-industrial complex'. In 2001 expenditure on military research and development was: (in millions of dollars) USA 39,340; (total EU 9,100;) Britain 3,986; France 3,145; Germany 1,286; Italy 291; Spain 174; Canada 121; Netherlands 65; Turkey 50. (Economist, 3/5/03). The total US expenditure on defence is about 340,000 per annum.

[41]M. L. Sifry and C. Cerf, “The Iraq War Reader: History Documents, Opinions, Simon and Schuster, New York, 2003.

[42]As I write (in London in July 2003) the crisis over the death of the principal British scientific expert on biological warfare, who found that the scientific situation was misrepresented by political leaders, exhibits the problems of the relationship between science and politicians. “What is truth said jesting Pilate, and would not wait for an answer” Francis Bacon (1561-1626).

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.

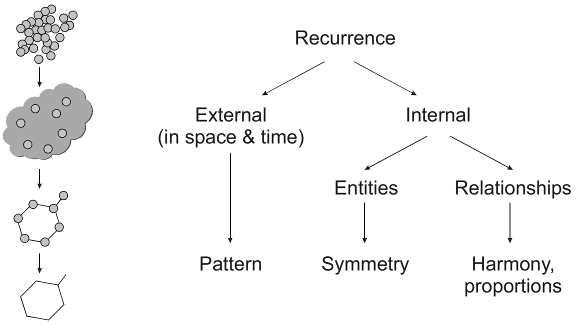


Figure 1. Simplified overview of concepts underlying structural descriptions.

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.

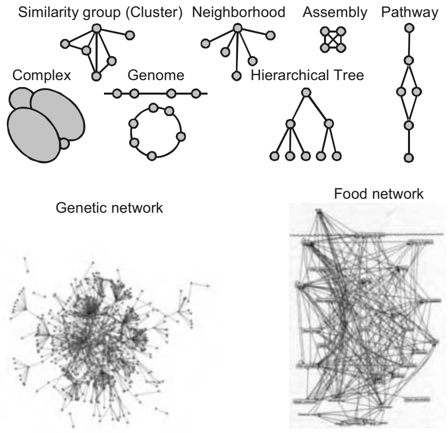


Figure 2. Molecular structures can be represented as entities and relationships [1, 13]. Implicit to a structure is the description of the underlying concepts (entities and relationships as well as their properties), which can be summarized in an ontology [14]. The same principle can be easily extended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

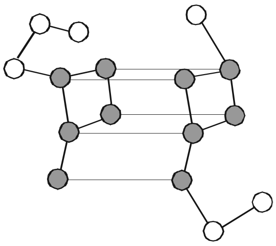


Figure 3. Molecular similarity as sub-graph isomorphism. Similarity of structures can be defined as a common sub-graph shared by two entity-relationship descriptions.

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.).

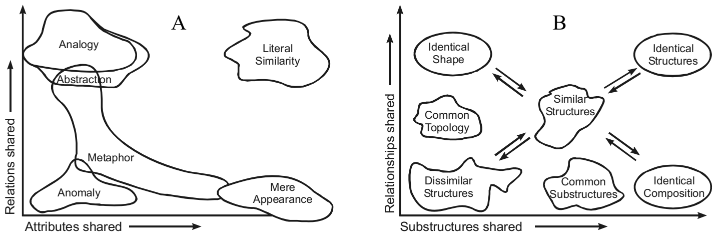


Figure 4. Identity, different kinds of similarity and non/identity can be pictured as regions in a plot of shared entities vs. shared relationships. This representation was developed by Dedré Gentner for narrative descriptions [17] (A), but can be extended to molecular descriptions as well (B).

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 ≅ A for all molecular descriptions A. It is symmetric if A ≅ B implies B ≅ A. It is transitive if A ≅ B, and B ≅ C implies A ≅ 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.

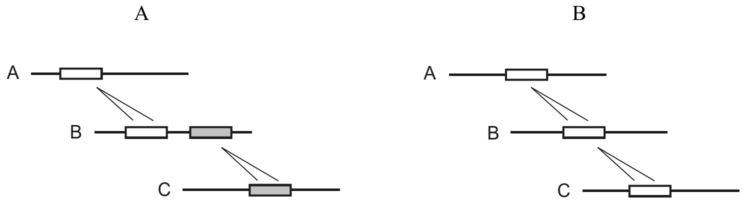


Figure 5. Similarity of molecules can be considered either a tolerance relationship (A), or an equivalence relationship (B) depending on whether or not the basis of similarity – the shared substructure – is fixed.

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 = ∞ 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]
$$S_{1,2} = \sum \cos t_{identities,replacementss} - \sum \cos t_{gaps}$$