Data Learning: Understanding Biological Data

Vladimir Brusic[1,2,5], John S. Wilkins[3], Clement A. Stanyon[2] and John Zeleznikow[4]

[1]Kent Ridge Digital Labs, Singapore.
{vladimir@krdl.org.sg }
[2]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.
{vladimir@wehi.edu.au; stanyon@wehi.edu.au }
[3]Department of the History and Philosophy of Science, The University of Melbourne, Melbourne, Victoria, Australia
{wilkins@wehi.edu.au}
[4]School of Computer Science and Computer Engineering, La Trobe University, Bundoora, Victoria, Australia.
{johnz@latcs1.cs.latrobe.edu.au}

Abstract

The four most important data-related considerations for the bioinformatic analysis of biological systems are understanding of: the complexity and hierarchical nature of processes that generate biological data, fuzziness of biological data, biases and potential misconceptions in data, and the effects of noise and errors. We discuss these issues and summarize our findings by defining a *Data Learning Process* (DLP). DLP comprises a series of steps for comprehension of biological data within the bioinformatics framework. DLP is a formalization aimed at facilitating knowledge discovery in biological databases.

Introduction

Biological databases continue to grow rapidly. This growth is reflected in increases in both the size and complexity of individual databases as well as in the proliferation of new databases. We have ever-increasing requirements for both speed and sophistication of data analysis to maintain the ability to effectively use the available data. Bioinformatics is a field emerging at the overlap between biology and computer science. Biological science provides deep understanding of this complex domain, while computer science provides effective means to store and analyse volumes of complex data. Combining the two fields gives the potential for great strides in understanding biological systems and increasing the effectiveness of biological research. The difficulties in effective use of bioinformatic tools arise at both ends: an average biologist has a limited understanding of sophisticated data analysis methods, of their applicability and limitations, while an average computer scientist lacks understanding of the depth and complexity of biological data. Bioinformaticians need to develop an overlap of understanding between the two fields. Here we discuss the issues related to biological data which have implications on selection and critical usage of computer science methods in biological research. The aim of this article is to clarify some important aspects of biological data for computer scientists.

5. corresponding author, V. Brusic, Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613

What do we need to know about biological data?

The four most important data-related considerations for the analysis of biological systems are understanding of: a) the complexity and hierar-

chical nature of processes that generate biological data, b) fuzziness of biological data, c) biases and potential misconceptions arising from domain history, reasoning with limited knowledge, a changing domain and methodological artefacts, and d) the effects of noise and errors. Despite a broad awareness, biological-data-specific issues have not been reported extensively in the bioinformatics literature. This awareness is exemplified in the words of Altschul *et al.*,1994:

> "Surprisingly strong biases exist in protein and nucleic acid sequences and sequence databases. Many of these reflect fundamental mosaic sequence properties that are of considerable biological interest in themselves, such as segments of low compositional complexity or short-period repeats. Databases also contain some very large families of related domains, motifs or repeated sequences, in some cases with hundreds of members. In other cases there has been a historical bias in the molecules that have been chosen for sequencing. In practice, unless special measures are taken, these biases commonly confound database search methods and interfere with the discovery of interesting new sequence similarities."

Biological data are sets of facts stored in databases which represent measurements or observations of complex biological systems. The underlying biological processes are highly interconnected and hierarchical; this complexity is usually not encoded in the data structure, but is a part of "background" knowledge. Knowledge of the biological process from which data are derived enables us to understand the domain features that are not contained in the data set. Raw information thus has a meaning only in the broader context, understanding of which is a prerequisite for asking "right" questions and subsequent selection of the appropriate analysis tools. According to Benton, 1996, the complexity of biological data is due both to the inherent diversity and complexity of the subject matter, and to the sociology of biology.

Biological data are quantified using a variety of direct or indirect experimental methods. Even in a study of a clearly delineated biological phenomenon a variety of experimental methods are usually available. An experimental method is considered useful if a correlation can be established between its results and a studied phenomenon. This correlation is rarely, if ever, perfect. Distinct experimental methods in the study of the same biological phenomenon would generally produce sets of results that overlap, but not fully. Comparing these results involves scaling and granularity issues. Within the same experimental method, differences of results arise from our inability to reproduce identical conditions (eg. temperature, pH, use of different cells or cell lines, use of chemicals from different suppliers etc.). Quantification of the results is commonly a result of a human decision or it may vary due to calibration of equipment. A reported quantitative result is typically the average value of several independent experiments. Quantitative biological data are fuzzy due both to inherent fuzziness of the biological systems themselves, and to the imprecision of the methods used to collect and evaluate data. Quantitative biological data therefore represent approximate measurements. On the other hand, the classes to which qualitative biological data are assigned are arbitrary, but objective in that they represent some biological fact. Biological research is largely driven by geographically dispersed individuals, who use unique experimental protocols and biological experimental data are produced with neither standard semantics nor syntax (Benton, 1996). Understanding the fuzzy nature of biological data is therefore crucial for the selection of appropriate data analysis tools.

A set of biological data rarely represents a random sample from the solution space. Typically, new results are generated around previously determined data points. Some regions of the solution space are therefore explored in depth, while some regions remain totally unexplored. Historical reasons are a common cause of such biases, where a set of rules might be defined in an attempt to describe a biological system. If these rules get accepted by a research community, further research will get directed by applying these rules. If those rules circumscribe a subset of the solution space, the consequence is the refinement of the knowledge of the subset of solutions that satisfies the rules, while the rest of the solution space remain largely

ignored. Similarly, reasoning with limited knowledge can lead to over- or under-simplification errors. A careful assessment of the relative importance of each data point is thus necessary for the data analysis. Improvements in the technology also influence biological data. Older data are often of lower granularity both quantitatively and qualitatively, while newer data are typically of higher precision, due to both expanded background knowledge and improved experimental technology.

Sources of noise in biological data include experimental, measurement, reporting, annotation and data processing errors. While it is not possible to eliminate errors from data sets, a good estimate of the level of noise within the data helps selection of the appropriate method of data analysis. Due to the complexity of biological systems, theoretical estimation of error levels in the data sets is difficult. It is often possible to make a fair estimate of the error level in biological data by interviewing experimental biologists who understand both the process that generated that data and the experimental methodology. In the absence of a better estimate it is reasonable to assume the error level in biological data at 5%.

To illustrate the above points we give an example where the usefulness of the overlap of biology and computer science has been demonstrated. Here we briefly describe the data learning process in bioinformatic prediction of patterns within peptides which can trigger and regulate immune responses.

Prediction of T-cell epitopes

The biology. T cells of the immune system in vertebrates recognise short antigenic peptides derived from the degradation of proteins. These peptides are presented on the surface of antigen presenting cells to the T cells by MHC (major histocompatibility complex) molecules. A cancer cell or a cell infected by a virus, for example, presents a subset of peptides that are different from those presented by a healthy cell. In a healthy organism, cells displaying 'foreign' antigenic peptides are destroyed by the immune response following T-cell recognition. Antigenic peptides therefore act as recognition labels for the immune system and are keys in the mechanism of triggering and regulation of the immune response. Antigenic peptides that mediate an immune reaction are termed T-cell epitopes. The ability to determine T-cell epitopes is critical for our understanding how the immune system functions and opens ways towards the design of peptidic drugs and vaccines.

The MHC/peptide binding problem. MHC molecules play a central role in immune interactions at the molecular level. Binding of a peptide to a MHC molecule is mediated through hydrogen bonds between the groove of the MHC molecule and the peptide backbone as well as through interaction between side chains of amino acids that form a peptide and specific pockets within the groove (Bjorkman *et al.,* 1987; Brown *et al.,* 1993). Peptide/MHC binding is thus influenced by its overall structure and by the side chains of the individual amino acids. Contribution of individual amino acids in particular positions within peptides may have positive, neutral or negative contribution to MHC binding. These contributions have been exemplified in binding motifs (Rammensee *et al.,* 1995). Binding motifs provide a qualitative description of the contribution to binding of each amino acid (of the possible 20) at a particular position within MHC-binding peptides. More than 500 variants of MHC molecules are known in humans (see Travers P.J. – *Histo* database). Different MHC molecules bind peptide sets that may be distinct or may overlap to various degrees. Prediction of T-cell epitopes is therefore possible only relative to specific MHC alleles. Furthermore, peptide binding to the MHC molecule is a necessary, but not sufficient condition for its 'T-cell epitopicity'. To be a T-cell epitope, a peptide must be recognised by a matching T cell and thus the T-cell epitopicity of a peptide can only be determined in the context of a target biological system (an organism or a particular cell line). The prediction of T-cell epitopes is often confused with the prediction of MHC-binding peptides. In determination of T-cell epitopes, prediction of MHC binding

peptides equates to the narrowing of the pool of potential T-cell epitopes.

The models. Three types of models that incorporate biological knowledge have been used for prediction of MHC binding peptides: *binding motifs* (Rammensee *et al.*, 1995), *quantitative matrices* (Parker *et al.*, 1994; Hammer *et al.*, 1994) and *artificial neural networks* (Brusic *et al.*, 1994; Brusic *et al.*, 1998a). Binding motifs (Fig. 1a) are the simplest models, which represent the anchoring patterns and the amino acids commonly observed at anchor positions. Quantitative matrices (Fig. 1b) provide coefficients that quantify contribution of each amino acid at each position within a peptide to MHC/peptide binding. Matrices encode higher complexity than binding motifs but ignore the effect of the overall structure of peptide, such as influences of neighbouring amino acids. We can encode an arbitrary level of complexity in artificial neural network (ANN) models (Fig. 1c) by varying the number of hidden layer nodes or the number of hidden layers. ANN models can therefore encode both the effects of the overall peptide structure and of individual amino acids to MHC/peptide binding. If sufficient data are available, more complex models perform better, as shown in a comparative study (Brusic *et al.*, 1998a). On the other hand, it is not beneficial to use models whose complexity exceeds the complexity of the process that generated data. This will increase required amounts of data for model building and possibly worsen the predictive performance of the model.

The data and analysis. The purpose of predictive models of MHC/peptide interactions is to help determine peptides that can bind MHC molecules and therefore are potential targets for immune recognition *in vivo*. Various experimental methods have been developed to measure (directly or indirectly) peptide binding to MHC molecules. Van Elsas *et al.*, 1996 reported the results of three experimental binding methods in determining T-cell epitopes in a tumour-related antigen (Melan-A/MART-1) in context of human MHC molecule HLA-A*0201. The summary of their report is given in Fig. 2, being an instance of poor correlation of results between various experimental binding methods. In the development of predictive models, we want to maximally utilize available data. Combining data from multiple experimental methods requires dealing with imprecise and inexact measurements. For MHC binding, fuzzy measures of high-, moderate-, low- and zero-affinity binding have been commonly used. The application of fuzzy logic (Zadeh, 1965) enables quantification of fuzzy data sets and the extraction of rules for model building. Artificial neural networks are particularly useful for extracting rules from fuzzy data (Kosko, 1993) and have been successfully used for prediction of MHC binding peptides (reviewed in Brusic and Harrison, 1998). By trimming ANN models of MHC/peptide binding we can demonstrate that binding motifs and quantitative matrices represent different levels of complexity of the same model, showing that the basic rules of MHC/peptide interactions, ie. background knowledge, has been captured in all these models.

Misconceptions and biases. A decade after the basic function of MHC molecules was described (Doherty and Zinkernagel, 1975), a small database of T-cell epitopes was compiled, followed by propositions of predictive models of T-cell epitopes. One such model (DeLisi and Berzofsky, 1985) was based on the assumption that a T-cell epitope forms an amphipatic helix[6] which binds into the groove of MHC molecules. Although the amphipatic model was incorrect, it was used for a decade. Those predictions that were fortuitously correct were also preferentially reported in the literature, reinforcing the presumed usefulness of the model. It was another decade before the models based on detailed knowledge of peptide/MHC interactions emerged (reviewed in Brusic and Harrison, 1998). Biases in available data arise from a non-critical usage of proposed binding motifs which reinforces data around peptides that conform well with proposed binding motifs. There are many examples of peptides that do not conform to the proposed binding motifs, yet bind the corresponding MHC

6. amphipatic helix – a helical structure of a peptide which has one side hydrophylic (attracts water molecules) and the other hydrophobic (repels water molecules).

A) A Binding Motif of human HLA-DRB1*0401 (Rammensee *et al*., 1995)

	Relative position								
	1	2	3	4	5	6	7	8	9
Anchor (bold), preferred or forbidden (italic) residues	**F,Y** **W**,I L,V M			F,W I,L D,E no *R,K*		N,S T,Q H,R	pol.* chg.* ali.*		pol.* ali.* K

*pol.: polar; chg.:charged; ali.:aliphatic residues.

B) A Quantitative Matrix of human HLA-DRB1*0401 (adapted from Hammer*et al*., 1994)

Position	Amino acid																		
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	*	*	*	*	0	*	*	−10	*	−10	−10	*	*	*	*	*	*	−10	0	0
P2	0	0	−13	1	8	5	8	11	11	10	11	8	−5	12	22	−3	0	21	−1	9
P3	0	0	−13	−12	8	2	2	15	0	10	14	5	3	0	7	2	0	5	0	8
P4	0	0	17	8	−8	−15	8	8	−22	−6	14	5	−21	11	−15	11	8	5	−12	−10
P5	0	0	−2	−1	3	2	−1	1	3	1	3	2	5	1	0	4	6	4	−1	−2
P6	0	0	0	−12	−13	−11	−16	−2	−23	−13	−13	17	1	−12	−22	17	19	13	−9	−11
P7	0	0	−11	−2	−8	−15	−8	−2	−12	4	7	−1	−3	−5	−12	−4	−2	5	−13	−7
P8	0	0	−11	−2	1	−5	0	−1	9	6	4	7	−2	16	7	6	5	4	6	13
P9	0	0	−25	−18	−8	−2	3	−4	−9	−13	−4	−11	−16	7	−9	12	−3	5	−3	−15

* forbidden amino acid

C) An ANN model for prediction of MHC-binding peptides (see Brusic *et al*., 1998a)

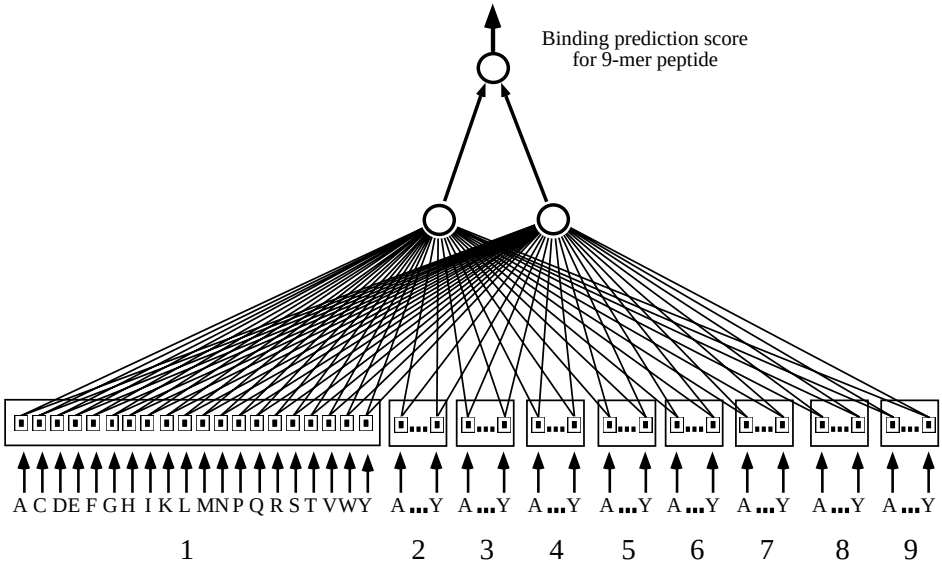

Figure 1. The models used for prediction of MHC-binding peptides. A) An example of a binding motif which indicates the positions and amino acids of main anchors, preferred and forbidden residues. B) A matrix that quantifies a contribution to MHC/peptide binding of each amino acid at each position of a 9-mer peptide. The predicted binding affinity is calculated as a sum of coefficients for amino acids within a peptide. C) An ANN model used to learn MHC-binding patterns which comprises 180 input layer units, 2 hidden layer units and a single output unit. A representation of an individual amino acid is a binary vector of length 20.

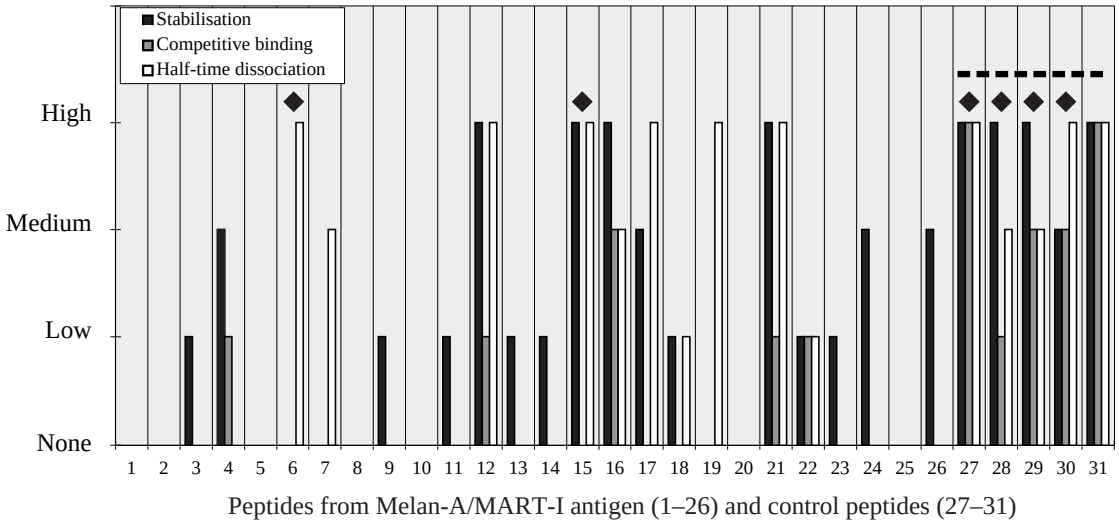

Figure 2. A summary comparison of the results three experimental methods for determination of HLA-A*0201 binding peptides from a tumour antigen MART-1 (adapted from van Elsas *et al.*, 1996). The fuzzy measures of binding affinity (high, moderate, low and none) are used at the vertical scale. Binding results for controls correlate well, while those for MART-1 peptides correlate poorly. The diamonds indicate T-cell epitopes.

molecule; many of these peptides are also reported as T-cell epitopes (see Brusic *et al.*, 1998b).

Sets of MHC-related peptides usually contain large subsets which comprise variants of a single peptide – single point mutation products. The information contained in these variant peptides is important for building accurate models, but it also introduces solution space biases. The first step in model building involves peptide alignment, and an uncritical usage of the complete set is likely to produce alignment skewed towards the over-represented subsets. An example of a de-biasing technique is given in Brusic *et al.*, 1998a where a scheme was used to weight peptides and penalise similarity between them.

Data errors. Noise and errors in the data sets affect our ability to derive useful models. Brusic *et al.*, 1997 studied the effect of noise in data sets on development of quantitative matrix models. They showed that the moderate level of noise significantly affects our ability to develop matrix models. For example, 5% of erroneous data in a data set will double the number of data points, relative to a 'clean' data set, required to build a matrix model of a pre-set accuracy. On the other hand, 5% of errors does not significantly affect the overall sucess of prediction of ANN models due to their ability to handle imperfect or incomplete data (Hammerstrom, 1993).

Conclusions

When sufficient data are available and the biological problem is well-defined, standard statistical methodology should be applied. A field where this approach has been routinely used is epidemiology (see Coggon *et al.*, 1997). Most of biological research, particularly in molecular biology, is conducted in domains characterized by limited background knowledge and by data from various sources and of variable accuracy. In such cases the artificial intelligence techniques are more useful, as shown in 'T-cell epitope prediction' example. Here we provide a set of guidelines which should help computer scientists to understand biological data and aid the design of the appropriate data analysis methods. To facilitate bioinformatic analysis of biological systems, we have defined a *Data Learning Process* (DLP), comprised of a series of steps (Fig. 3). The DLP steps are: a) develop understanding of the biological system and methodological processes that generate data, b) develop a

standardized fuzzy representation of the data, c) relate data from various sources using this standardized representation, d) identify potential sources of biases in data, e) assess the validity of relevant models reported in the literature, f) estimate the amount and types of errors in the data sets, and g) integrate knowledge acquired in previous steps in some coherent form (e.g., model or description). Performing the DLP steps requires significant inputs from both biologists and computer scientists and must involve two-way communication.

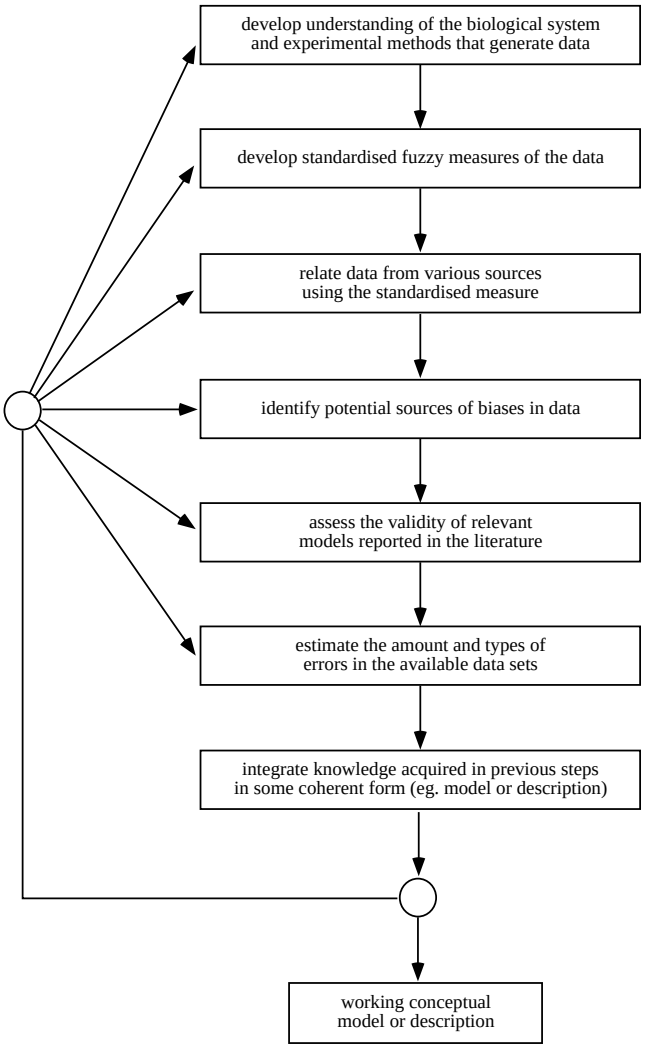

Figure 3. A flow diagram of *Data Learning Process.*

A useful starting point is the development of a conceptual model of the studied biological system. Conceptual graphs (Sowa, 1984) integrate formal logic and the clarity of graphic representation, thus providing notation that is understandable to experts from different fields and which is useful for formalization and unification of interdisciplinary knowledge. The DLP is an iterative process in which all of the latter steps help improve the first step – understanding of the background biology and methods that generate data. The model or description generated by DLP can then be used as a starting point for the design of a data analysis algorithm and the selection of the appropriate data analysis tools. Learning the application domain is the first step in the process of *Knowledge Discovery in Databases* (Fayyad *et al.*, 1996). DLP is thus a formalization of the procedure aimed at the facilitation of knowledge discovery in biological databases.

References

Altschul S.F., Boguski M.S., Gish W. and Wootton J.C. (1994). Issues in searching molecular sequence databases. *Nature Genetics*. 6(2):119–129.

Benton D. (1996). Bioinformatics – principles and potential of a new multidisciplinary tool. *Trends in Biotechnology* 14:261–272.

Bjorkman P.J., Saper M.A., Samraoui B., Bennett W.S. and Strominger J.L. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329(6139):506–512.

Brown J.H., Jardetzky T.S., Gorga J.C., Stern L.J., Urban R.G., Strominger J.L. and Wiley D.C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364(6432):33–39.

Brusic V. and Harrison L.C. (1998). Computational methods in prediction of MHC-binding peptides. In Michalewicz M. (ed.), *Advances in Computational Life Sciences: Humans to Proteins*, p. 213–222, CSIRO Publishing, Melbourne.

Brusic V., Rudy G. and Harrison L.C. (1994). Prediction of MHC binding peptides using artificial neural networks. In Stonier R.J. and Yu X.S., (eds*), Complex Systems: Mechanism of Adaptation*, pp. 253–260, IOS Press, Amsterdam/ OMSHA Tokyo.

Brusic V., Rudy G., Honeyman M.C., Hammer J. and Harrison L.C. (1998a). Prediction of MHC class-II binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 14(2):121–130.

Brusic V., Rudy G. and Harrison L.C. (1998b). MHCPEP – a database of MHC-binding peptides: update 1997. *Nucleic Acids Research* 26:368–371.

Brusic V., Schoenbach C., Takiguchi M., Ciesielski V. and Harrison L.C. (1997). Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. *ISMB* 5:75–83.

Coggon D., Rose G. and Barker D.J.P. (1997). *Epidemiology for the Uninitiated*. Fourth edition. BMJ Publishing Group, <http://www.bmj.com/epidem/epid.html>

DeLisi C. and Berzofsky J.A. (1985). T-cell antigenic sites tend to be amphipathic structures. *Proceedings of the National Academy of Sciences of the United States of America* 82(20):7048–7052.

Doherty P.C. and Zinkernagel R.M. (1975). A biological role for the major histocompatibility antigens. *Lancet* 1(7922):1406–1409.

Fayyad U., Piatetsky-Shapiro G. and Smyth P. (1996). From data mining to knowledge discovery. *AI Magazine* 17(3):37–54.

Hammer J., Bono E., Gallazzi F., Belunis C., Nagy Z. and Sinigaglia F. (1994). Precise prediction of MHC class II-peptide interaction based on peptide side chain scanning. *Journal of Experimental Medicine* 180:2353–2358.

Hammerstrom D. (1993) Neural networks at work. *IEEE Spectrum* 30:26–32.

Kosko B. (1993). *Fuzzy Thinking. The New Science of Fuzzy Logic*. Harper Collins Publishers, Glasgow.

Parker K.C., Bednarek M.A. and Coligan J.E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of Immunology* 152:163–175.

Rammensee H.G., Friede T. and Stevanovic S. (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics* 41:178–228.

Sowa J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley Publishing Co. Inc., Reading, Massachusetts.

Travers P.J. *Histo.* <http://histo.cryst.bbk.ac.uk//>

van Elsas A., van der Burg S.H., van der Minne C.E., Borghi M., Mourer J.S., Melief C.J. and Schrier P.I. (1996). Peptide-pulsed dendritic cells induce tumoricidal cytotoxic T lymphocytes from healthy donors against stably HLA-A*0201-binding peptides from the Melan-A/MART-1 self antigen. *European Journal of Immunology* 26(8):1683–1689.

Zadeh L.A. (1965). Fuzzy Sets. *Information and Control* 8:338–353.