VLADIMIR BRUŠIĆ

Institute for Infocomm Research, Singapore

COMPUTATIONAL MODELS FOR MOLECULAR MEDICINE

Molecular medicine focuses on diagnosis, treatment, and prevention of injury and disease by targeting defects at the molecular level. Diagnostic tests, therapeutics, or prophylactics derived from DNA or protein sequence information represent major targets in molecular medicine research. Research in this field is largely experimental, but systematic experimental studies are often prohibitively expensive or even not possible. Computational modeling has emerged as a convenient technology in support of experimental research. It is particularly useful in situations when combinatorial nature of the studied problem requires thousands or even millions of individual experiments. Preliminary screening using accurate models of molecular interactions can identify a small number of key experiments that are sufficient for completing the study.

High throughput technologies such as genomics or proteomics produce large amounts of data of gene and protein expression (Auffray et al., 2003). The combinatorial nature of molecular processes, makes combination of computational modeling and experimentation necessary for molecular medicine studies, such as those involving transcriptional regulation (Beer and Tavazoie, 2004), or antigen processing and presentation (Flower, 2003). Computational models are useful in biological research for selection and planning of experiments as well as for interpretation and understanding the implications of accumulated data (Brusic and Zeleznikow, 1999).

In this article we discuss the requirements for the development and use of computational models and demonstrate the evolution of computational modeling with the accumulation of knowledge in a given domain. This will be demonstrated using examples from identification of immune epitopes that represent targets for vaccine development.

Author address: V. Brusić, Institute for Infocomm Research, Singapore

Paper presented as a lecture.

REQUIREMENTS FOR BUILDING COMPUTATIONAL MODELS

For maximum benefit, computer models must be developed and assessed for performance before use; this assessment must be done with the same rigor as it is done with standard laboratory procedures. This involves due care in testing and validation of computational models, design of simulated experiments, and interpretation of results. The best modeling and computational practices must be applied to the design and development of computational models. Before use, computational models must be assessed for relevance, accuracy, generalization properties, precision, and robustness. Computer models that are relevant and accurate can be used to complement laboratory experimentation and have been termed computational assays (Brusic and Zeleznikow 1999). Model relevance refers to the correctness of the assumptions – they must be in line with current scientific and technical knowledge related to the studied problem. For example, models that predict signal peptides cannot be used for identification of transmembrane regions of proteins. Several measures of accuracy of prediction models are known (Bajic, 2000). The commonly used measures are shown in Table 1.

Generalization properties indicate the ability of a model to accurately predict new cases and therefore its practical utility. Poor generalization ability may arise from a) inadequate data sets used for training, b) inadequate selection and use of learning algorithm, or c) mismatch of the complexity of model with the that of either the modeled phenomena or the amount of training data. When data used for model building are overspecialized, i.e. represent only a subset of the relevant data, the model may not be adequate for the intended use. For example, protein structure models derived for globular peptides will perform poorly if used for prediction of structure of transmembrane domains of proteins. Poor generalization may result from model over–training – a situation when a model is forced to learn the peculiarities of the training data set, rather than general rules. Finally, overly complex models tend to memorize training examples, but will often not learn general rules.

An overly simple model will have a limited ability to learn, resulting in a lower accuracy of predictions.

Table 1. Definition of terms for assessment of the accuracy of predictive models.

	Experimental positives	Experimental negatives
Predicted positives	True positives (TP)	False positives (FP)
Predicted negatives	False negatives (FN)	True negatives (TN)

Accuracy measure	Formula	Pairs with
Sensitivity	SE = TP/(TP + FN)	SP
Specificity	SP = TN/(TN + FP)	SE
Positive predictive value	PPV = TP/(TP + FP)	NPV
Negative predictive value	NPV = TN/(TN + FN)	PPV
Accuracy	Acc = (TP + TN)/ /(TP + TN + FP + FN)	–
Aroc	Integration of ROC curves (Swets, 1988)	–

CASE STUDY: PREDICTION OF IMMUNE EPITOPES

T cells of the immune system recognise short peptides, bound by major histocompatibility complex (MHC) molecules and displayed on the surface of host cells. These peptides are recognition labels, which display contents of host cells to T cells of the immune system. The presence of non–homeostatic (also known as non–self) peptides is a prerequisite for the initiation of immune responses. Peptides produced by degradation of intracellular proteins bind MHC class I molecules. MHC class II molecules present peptides on antigen–presenting cells produced by degradation proteins of extracellular origin. A major function of cytotoxic T cells is to recognise (by T–cell receptors) and destroy infected (e.g. viruses, bacteria), mutated (e.g. tumor) or foreign (e.g. transplant) cells. The availability of intracellular proteins and processing pathways determine a) which peptides are available for presentation by MHC–class–I pathway, and b) the extent of subsequent cytotoxic response. Peptides displayed by MHC class II molecules mainly serve to regulate immune responses; they are crucial for the initiation, enhancement and suppression of immune responses. The peptide–binding site of MHC molecules is a cleft comprising b–sheet supporting a pair of a–helices. A peptide binds through a network of hydrogen bonds between the peptide backbone and the cleft, as well as through interactions between peptide side chains and specific pockets in the cleft (Madden et al., 1993). Interaction between peptide and the binding cleft is mainly through primary and secondary anchors, which are the positions within peptide that provide strongest contribution to binding. For a given MHC molecule only a limited set of amino acids can act as anchors at a particular position within a peptide. Anchor positions are the key for defining common patterns within sets of peptides that bind a specific MHC molecule. Binding motifs to more than 200 MHC specificities have been proposed (Rammensee et al. 1999). These motifs provide a basis for the development of methods for predicting peptide binding to MHC molecules. An example of a binding motif is shown in Table 2. Binding motifs represent the most basic models of peptide binding to MHC molecules since they indicate approximate preference for certain amino acids at certain positions in peptides that bind a given MHC molecule. Binding motifs have the lowest accuracy of reported MHC–binding prediction methods (Yu et al., 2002).

Table 2. Binding motif for mouse MHC molecule H–2–Kb. (Rammensee et al., 1999).

	Position								
	1	2	3	4	5	6	7	8	9
Anchors					F				L
					Y				M
									I
									V
Auxiliary anchors			Y						
Preferred residues	R	N	P	R		T	N		
	I			D		I	Q		
	L			E		E	K		
	S			K		S			
	A			T					

Quantitative matrices are refined binding motifs derived from experimental data. This refinement requires assessment of contribution to binding of each amino acid at each position in peptide. Such matrices have been derived using experimental data. Summing up the coefficients for amino acids at each position in a peptide produces a binding score and scores above a specified threshold represent predicted binders. The matrix for HLA–DR4 (Hammer et al., 1994) correctly predicted > 70% of both binding and non–binding peptides. Quantitative matrices are efficient, easy to use and are more accurate than binding motifs. The interaction between peptides and MHC molecules are non–linear (Yu et al., 2002) while matrices and motifs only describe linear relationships. ANNs are more complex than motif– or matrix–based models and they require more binding data for training and the data pre–processing. Pre–processing involves peptide alignment and conversion to the format acceptable by the ANN software. Simple motifs and binding matrices can aid the peptide pre–processing step. For example, the

information of primary anchors was used for elimination of false positives in prediction of peptide binding to human class II molecule HLA–DR4 (Brusic et al., 1998). Hidden Markov models (HMM) use the probabilistic framework to map the search space onto a set of states. They can learn generalised probabilistic rules from data sets. HMMs were applied in prediction of HLA–A2 binding peptides (Mamitsuka, 1998). Another type of sophisticated classification models, support vector machines (SVM), have also been used for study of MHC–binding peptides (Zhao et al., 2003). HMMs and ANNs have been combined for prediction of promiscuous MHC–binding peptides and immunological hot–spots inside antigens (Srinivasan et al., 2004). Molecular modelling encompasses detailed knowledge of the crystal structure of MHC molecules and of protein–peptide interactions and has been used for prediction of peptide binding to MHC molecules (Schafroth and Floudas, 2004). The accuracy of 3D molecular models needs to be improved before they can be used for large–scale new predictions.

CONCLUSION

Computational models are important complementary methodology to experimental research. They are particularly useful in fields requiring large number of experiments due to the combinatorial nature of underlying systems and processes. In such fields, computational models evolve with the amount of data and the accumulation of knowledge. We described the evolution of computational models used for study of immune epitopes. The early models, based on binding motifs, indicated rough regularities in the peptide data sets. The next generation, binding matrices, represent linear models that quantify each position in peptide. They have been superseded by sophisticated non–linear models using ANNs, HMMs, or SVMs. The models based on 3D analysis complement the set of data–driven models. Finally, individual models can be combined for identification of immunological hot–spots and promiscuous epitopes, that are the best targets for vaccine discovery.

BIBLIOGRAPHY

[1] Auffray C., et al., From functional genomics to systems biology: concepts and practices. C. R. Biol. **326** (10–11) (2003) 879–892.

[2] Bajic V.B., Comparing the success of different prediction software in sequence analysis: a review. Brief. Bioinform. **1**(3) (2000) 214–228.

[3] Beer M.A., and Tavazoie S., Predicting gene expression from sequence. Cell **117**(2), (2004) 185–198.

[4] Brusic V. and Zeleznikow J., Computational binding assays of antigenic peptides. Lett. Pept. Sci. **6** (1999) 313–324.

[5] Brusic V., et al., Prediction of MHC class II–binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics **14**(2) (1998) 121–130.

[6] Cai Y.D., et al., Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides **24**(1), 159–161.

[7] Flower D.R., Towards in silico prediction of immunogenic epitopes. Trends Immunol. **24**(12) (2003) 667–674.

[8] Madden D.R., et al., The antigenic identity of peptide–MHC complexes: a comparison of the conformations of five viral peptides presented by HLA–A2. Cell **75**(4) (1993) 693–708.

[9] Mamitsuka H., Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins **33**(4) (1998) 460–474.

[10] Rammensee H.G., et al., SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. **50**(3–4) (1999) 213–219

[11] Schafroth H.D. and Floudas C.A., Predicting peptide binding to MHC pockets via molecular modeling, implicit solvation, and global optimization. Proteins. **54**(3) (2004) 534–556.

[12] Srinivasan K.N., et al., Prediction of class I T–cell epitopes: evidence of presence of immunological hot spots inside antigens. Bioinformatics (2004) (in press).

[13] Swets J.A., Measuring the accuracy of diagnostic systems. Science **240**(4857) (1988) 1285–1293.

[14] Yu K., et al., Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol. Med. **8**(3) (2002) 137–148.

[15] Yuan Z., et al., SVMtm: support vector machines to predict transmembrane segments. J Comput. Chem. **25**(5) (2004) 632–636.

[16] Zhao Y., et al., Application of support vector machines for T–cell epitopes prediction. Bioinformatics **19**(15) (2003) 1978–1984.