Chapter XXVI
In Machina Systems for the Rational *De Novo* Peptide Design

Paul Wrede
Charité Universitaetsmedizin Berlin, Germany

ABSTRACT

Peptides fulfill many tasks in controlling and regulating cellular functions and are key molecules in systems biology. There is a great demand in science and industry for a fast search of innovative peptide structures. In this chapter we introduce a combination of a computer-based guided search of novel peptides in sequence space with their biological experimental validation. The computer-based search uses an evolutionary algorithm that includes artificial neural networks as fitness function and a mutation operator, called the PepHarvester. Optimization occurs during 100 iterations. This system, called DARWINIZER, is applied in the de novo design of neutralizing peptides against autoantibodies from DCM (dilatative cardiomyopathy) patients. Another approach is the optimization of peptide sequences by an ant colony optimization process. This biologically-oriented system identified several novel weak binding T-cell epitopes.

WHAT MEANS PEPTIDE DESIGN?

Peptides regulate and control many cellular processes. Many cell-cell interactions make use of peptide recognition and binding. Peptides serve as hormones like ACTH and vasopressin or intercellular signaling molecules producing only a specific response in target cells after interaction with the cognate receptor. Most receptors bind only a single or a group of closely related molecules. The humoral immune system synthesizes antibodies and the antigen is often a peptide. A successful application of neutralizing antibody binding by *de novo* designed peptides is described in detail (Schneider, Wrede 1993; Schneider et al. 1998). In contrast the cellular immune system works with peptides as mediators between antigen presenting cells and T-cells. The binding of peptides to the MHC I receptor and the T-cell receptor

depends on a variety of similar peptide sequences. In the beginning of the nineties a special binding motif described two anchor positions. In the position 2 and 9 a hydrophobic amino acid with a large side chain seems to be important (Rammensee et al. 1995; Lund et al. 2005). Since this pattern seems to be peculiar for MHC I binding peptides it was introduced into a prediction tool called SYFPEITHI (Rammensee 1999). But many recent studies revealed that this pattern is not sufficient for a prediction with high accuracy. Still all available MHC I binding peptide prediction tools have a disappointing reliability (Peters et al. 2006; Filter, Wrede unpublished observations).

Often amino acid sequence patterns are not unique although they fulfill the same function meaning occupying the same binding sites of the target molecule. This makes the development of prediction and design tools an extraordinary endeavour. Some solutions to overcome these hindrances are described in the next section.

Several other sources describe the combinatorial chemical process of peptide design. This knowledge is also necessary and included into the computer-based rational peptide predictions. But the article here focuses on the computer-based rational peptide design.

BIOINFORMATIC TOOLS FOR THE COMPUTER-AIDED MOLECULAR DESIGN

I highlight novel techniques for molecule design especially peptide design and molecular feature extraction, which can be applied when three-dimensional molecular structures are not available. A necessary prerequisite for any rational attempt to identify or even design molecules with a desired property or activity is an accurate model of the underlying sequence- (structure)-activity relationship (SAR). Such SAR models serve as guideline in the search for novel and optimized compounds in evolutionary design cycles which have become possible due to advances in both compound generation and screening technology. It is obvious that the quality of the model determines the success rate of this multi-dimensional design process. Only if a relevant SAR model is used a rational molecular design can be successful (Wrede, Schneider 1994; Schneider, Soo 2003; Wrede, Filter 2006).

How can we develop a good SAR model? It is apparent that no cure-all recipe exists, nevertheless some general rules of thumb can be given. One approach is to consider the task as a pattern recognition problem, where three main aspects must be considered: first the data used for generation of a SAR hypothesis should be representative of the particular problem; second the way molecular structures are described for model generation and its level of abstraction must allow for a reasonable solution for the pattern recognition task; and third the model must permit non-linear relationships to be formulated since the interdependence between molecular activities and structural entities is generally non-linear. The first point seems to be trivial but selection of representative data for hypothesis generation is very difficult and often impossible due to a lack of data. The focus here is on the two latter points, namely different levels of data representation and descriptor types, and non-linear feature extraction from a given data set by artificial neural networks (ANN). Various types of ANN are of considerble value for many fields of research, including chemistry, biology, medicine, and pharmaceutical research. Main tasks performed by these systems are:

- Feature extraction
- Function estimation and non-linear modeling