Journal of Computer-Aided Molecular Design, **12:** 543–556, 1998.
KLUWER/ESCOM

Patenting computer-designed peptides

Shail Patel, Ian P. Stott, Manmohan Bhakoo & Peter Elliott
Unilever Research, Port Sunlight Laboratory, Bebington, Wirral L63 3JW, U.K.

Received 15 November 1997; Accepted 27 March 1998

Key words: bactericidal peptides, fitness landscapes, genetic algorithm, molecular modelling, neural network, patent space, QSAR

Summary

The problem of designing new peptides that possess specific properties, such as bactericidal activity, is of wide interest. Recently, attention has focused on the use of Computer-Aided Molecular Design techniques in parallel with more traditional 'synthesise and test' methods. These techniques may typically use Genetic Algorithms to optimise molecules based on Neural Network models that predict activity. In this paper we describe a successful application of this Molecular Design methodology that has resulted in novel bactericidal peptides of real value. A key issue for commercial utilisation of such results is the ability to protect the intellectual property rights associated with the discovery of new molecules. Typically peptide patents use structural templates of amino acid hydrophobicity-hydrophilicity that define highly regular peptide patent spaces. In an extension of established patenting practice we describe a patent application that uses a Neural Net predictive model to define the regions of peptide space that we claim within the patent. This formalism makes no a priori assumptions about the regularity of the patent space. A preliminary comparative investigation of the shape and size of this and other bactericidal peptide patent spaces is conducted.

Introduction

Bactericidal peptides are of special interest for a wide range of industrial, pharmaceutical and medical purposes. Recently significant progress has been made in the use of Computer-Aided Molecular Design (CAMD) techniques for the design of novel molecules possessing desired properties, e.g. polymers with glass transition, resistivity and conductivity [1, 2] or mechanical and shrinkage [3] properties; peptides or proteins with folding [4–6], binding [7–9] or cleavage [10] properties. There are a number of good reviews of this general field [11–13].

The approach is best used in parallel with more traditional 'synthesise and test' methods and typically relies on two stages:

(1) Forward Modelling: The use of non-linear modelling methods such as Neural Networks (NN) to predict molecular properties, often called Quantitative Structure Activity Relationships (QSARs). These methods build predictive models based on experimental data. They may use molecular parameters derived from the 3-D structure of the peptide, or a structural description of the molecule [1, 14].

(2) Model Inversion / Optimisation: The use of optimisation algorithms to invert the QSAR models to find new molecules of high activity. The inversion is generally one-to-many, e.g. there are many molecules that have the same activity, and so the inversion is more reasonably treated as an optimisation problem in the space of all molecules of the particular class. Genetic Algorithms [15, 16] have been found to be a successful way of designing molecules in this way.

This paper reports on a successful application that has resulted in novel peptides with experimentally demonstrated bactericidal activity. The experimentation validates the forward modelling, and the CAMD methodology generates many 'virtual' peptides with high-predicted bactericidal activity. These peptides may have commercial value, and hence there is a need to protect the intellectual property rights associated with them. In an extension of established patenting

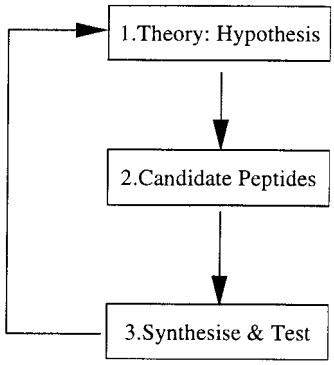

Figure 1. Primary Loop: Conventional synthesise and test methodology draws on a theoretically based hypothesis for the mechanism of action (1) to generate a set of peptides which may prove or disprove the hypothesis (2). A selection of these are synthesised and tested for activity (3).

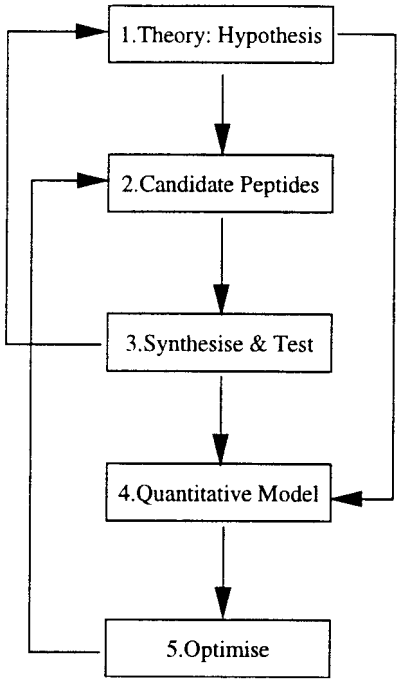

Figure 2. Secondary Loop: Descriptive molecular parameters based on the mechanism of action are selected or created, and together with data generated by the experimental synthesise and test loop are used to build a quantitative predictive model (4). Optimal peptides are generated using optimisation techniques (5), and a selection of these (2) are synthesised and tested (3).

practice we have framed a patent that uses the Neural Net QSAR model to define the regions of peptide space that we claim within the patent [17]. The Neural Net predictive QSAR model is not itself patented, it is fully disclosed, as are the methods of deriving the molecular parameters, so that anyone 'skilled in the art' may reproduce the results and models we have obtained. What the patent seeks to protect is the set of peptides obtained by application of CAMD techniques. This patent has a priority date of 9 March 1995 and was published by the European Patent Office on 19 September 1996.

Design cycle

A conventional 'synthesise and test' methodology starts with a hypothesis, generates a number of candidate peptides to test, conducts experiments on a selection of these peptides, and uses the results to confirm or falsify the hypothesis. Figure 1 shows this 'primary' loop. A CAMD approach puts in place a secondary loop of modelling and optimisation, where the two steps of forward modelling and model inversion form part of a total cycle of activity (Figure 2):

(1) Develop / Refine theoretically-based hypotheses for the mechanism(s) of actions.

(2) Generate candidate peptides (initially based on theory/hypothesis).

(3) Perform Experiments: synthesise and test a selection of candidate peptides.

(4) Develop a Quantitative Model: build a correlation model between peptide and activity, e.g. based on selected molecular parameters.

(5) Optimise Peptides using Model: generate candidate peptides based on quantitative model.

Once the primary loop has generated a sufficient quantity of data for modelling, the secondary modelling loop may commence, generating candidate peptides based on the experimental data. Note that the models and the results of the optimisation may also be used to influence the refinement of the theoretically based hypotheses. This process is similar to the hypothetico-deductive scientific process, and makes a compelling argument for placing computer aided design techniques firmly alongside traditional 'synthesise and test' methods.

Hypothesis: mechanism of action

Peptides that are capable of broad spectrum bactericidal activity should possess the ability to traverse a cell wall or outer cell wall membrane and disrupt or disintegrate cell membranes, in particular the cytoplasmic membrane which is a selective barrier that

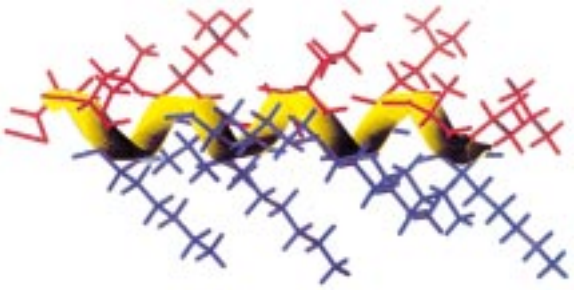

Figure 3. Side view of an amphiphilic and laterally amphipathic peptide with one face of the helix displaying hydrophobic residues, while the opposite face displays hydrophilic residues (red = hydrophobic, blue = hydrophilic).

restricts entry and exit of solutes. This results in irreversible osmotic-colloidal interactions which kill the bacteria by dramatically altering the proton motive force, i.e. $\Delta pH + \Delta\Psi$ (membrane potential). The peptides suitable for this purpose should be capable of forming ion channels in membranes by aggregation (pseudoionophores) and insertion in order to span the membranes [18–20]. The minimum length to span a membrane 2.5–4.0 nm (depending on lipid composition) is at least 15 amino acid residues [21, 20]. The peptides should be alpha-helical to be transmembrane [22–24]. They should also be amphiphilic and laterally amphipathic (one face of the helix displaying hydrophobic residues, while the opposite face displays hydrophilic residues), in order to form hydrophilic ion channels or pores and at the same time remain in contact with the hydrophobic components, e.g. fatty acyl moieties [25, 26] (Figures 3 and 4). A number of polar, acidic or basic amino acids are required within the peptide to impart suitable solubility characteristics [27].

Transmembrane pores or channels may arise from the insertion of monomeric peptides, where the pore is part of the secondary structure of the monomer, or from aggregated monomers which form oligomeric peptides, in which case a 'barrel-stave' type of pore is formed [28] (Figure 5). The 'barrel-stave' type of pores may arise by direct insertion of an oligomer into the membrane or by monomer insertion followed by lateral coalescence. Our peptides are designed as lateral amphipathic rods, which favour 'barrel-stave' type of pores.

These theories of mechanism of action form a set of 'design rules', which can be used to design peptides that should be active. The CAMD approach formalises the design rules in a quantitative model, that given a particular peptide will predict its activity with a high degree of accuracy. We have pursued both routes in parallel.

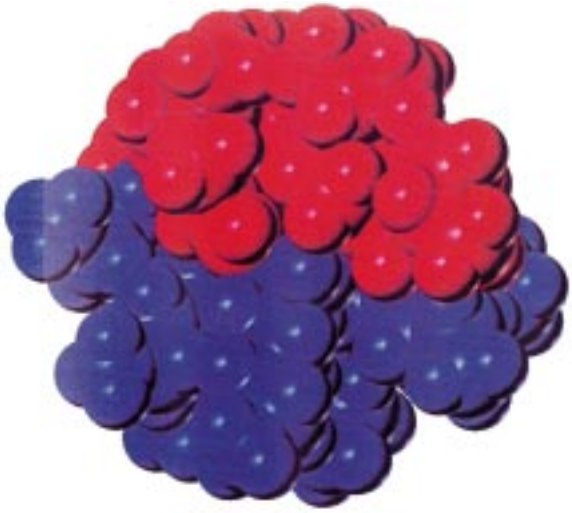

Figure 4. Space-filling view of the peptide in Figure 3, looking down the helical axis.

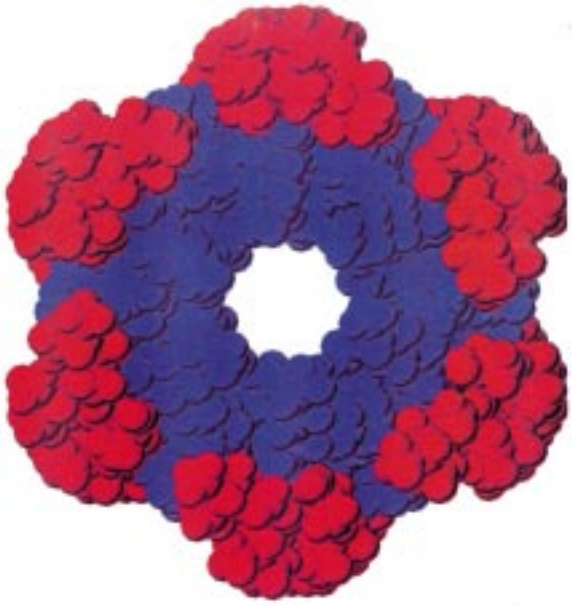

Figure 5. Space-filling view of an aggregation of six peptides showing a 'barrel-stave' pore which may form an ion channel in the cell membrane, orientation as in Figure 4.

Experimental measures

Test bacteria *Staphylococcus aureus* ATCC 6538 and *Escherichia coli* ATCC 11229 were cultured

overnight, harvested and suspended in Ringer's solution to obtain appropriate initial cell numbers (10^6–10^8). The bacteria were incubated with the designed test peptides (various concentrations) aseptically in 96 well microtitre plates and relevant controls taken. The bactericidal activity (bacterial killing) was measured by a total viable counts (TVC) technique after an incubation period of 2 h with the test peptides, and the results expressed as log bacterial kill. The log bacterial kill is calculated as the difference in log scale of bacteria initially present and the number of bacteria still viable after incubation with peptides.

Modelling techniques

Over the period of study, a total of 29 peptides were synthesised and tested for the bactericidal activity, and a further 5 were synthesised and tested based on the model predictions (cf. Appendix 1). We present here in detail the modelling of bactericidal activity against *S. aureus*, the case against *E. coli* runs in a similar vein. In this study, we adopt a two-stage approach to modelling: the first step is to generate a number of molecular descriptors based mainly on the 3-D structure of the peptides, and the second step is to find a relationship between these parameters and the bactericidal activity (Figure 6).

Molecular modelling

Based on the mechanism of action, 39 molecular parameters were selected or created (using the TRIPOS-Sybyl software [29]), that might decribe the key aspects of the peptides structure that are responsible for its activity against *S. aureus* and *E. coli* (cf. Appendix 2). Given an appropriate peptide length of greater than 15 amino acid residues, we assume:

$$\text{bactericidal activity} = \text{f(diffusion, aggregation)}$$

where the ability to diffuse and aggregate is described by two groups of parameters: general molecular parameters and amphipathic descriptors. General molecular parameters include molecular weight, charge, size and shape parameters. An example of an amphipathic parameter is the hydrophobic dipole moment, a property analogous to the standard electrostatic dipole moment but using hydrophobicity of the amino acids instead of atomic point charges. The *x*, *y*, and *z* components of the dipole are derived by orienting the peptides with *z* down the centre of the alpha helix of the peptide with origin ($z = 0$) at the midpoint; the *y*-axis defined as being in the direction of the vector sum over all hydrophobic residues $\Sigma H_i C_i$, where H_i is the hydrophobicity value of Eisenberg [22, 23] and C_i is the co-ordinates of the Cα atom of amino acid i. Electrostatic dipole moments are calculated in a similar manner, using all atoms within the peptide instead of just the Cα atoms.

For a peptide to be laterally amphipathic, hydrophobicity should be periodic along the sequence of amino acids, with period ∼3.6, the circumference of a turn in the alpha-helix. Two parameters were devised to approximate to this, the closeness of fit of hydrophobicity to a sine wave, s_{sin}, and the closeness of fit to a square wave, s_{sqr}. These parameters are given by 'least-squares' equations:

$$s_{sin} = \text{Min}_\theta\{\Sigma_i^N[H_i - \sin(100i - \theta)]^2\}/N,$$

$$s_{sqr} = \text{Min}_\theta\{\Sigma_i^N\{0.5 * [\text{sign}(H_i) - \text{sign}(\sin(100i - \theta))]^2\}\},$$

where N is the number of amino acids in the peptide, i is the position along the amino acid sequence, H_i is normalised such that the largest positive and largest negative value of hydrophobicity for the residues are set at 1 and −1 respectively, θ is an offset that is varied to find minimum value.

Neural networks

To build an effective NN model, we employed the following steps: initial data exploration (including linear modelling), parameter selection, choice of model architecture, model training with cross-validation and blind testing.

Of the 39 initial parameters a smaller subset was selected. The auto-correlation between the input parameters highlighted clusters of similar parameters. Together with the coefficients of the input parameters, calculated by a step-wise linear regression, six key variables were identified that made sense given the mechanism of action:

- Sum of the negative charge (1).
- Sum of the hydrophobic and hydrophilic values (2).
- closeness of fit to 'sine' wave and 'square' wave (3 and 4).
- *x* and *y* components of the charge dipole (5 and 6).

These represent a selection of general molecular and amphiphilic parameters, i.e. of diffusion and aggregation parameters.

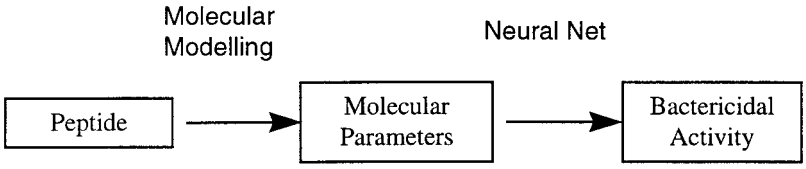

Figure 6. The first step of Molecular Modelling generates a number of molecular parameters based on the general characteristics and the 3-D structure of the peptide. These parameters are used as input to a Neural Network to give a prediction of the bactericidal activity.

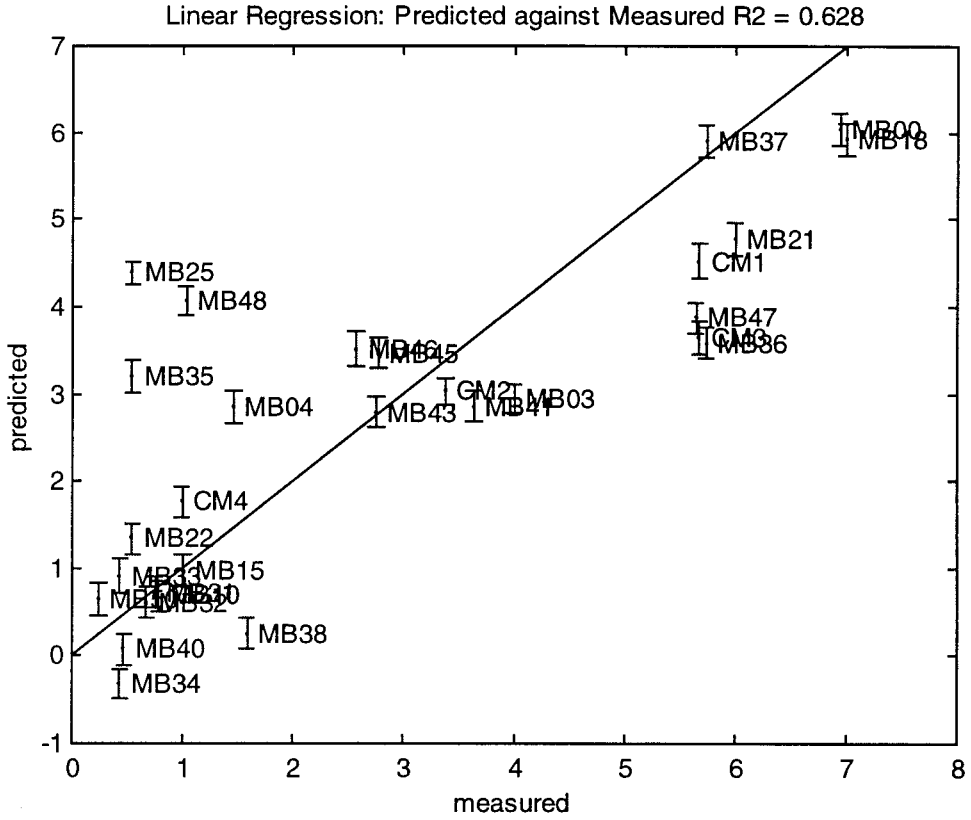

Figure 7. Results of linear regression, showing predicted log kill of *S. aureus* against measured values. The error bars calculated by the QR method are not good estimates of the true error. The R-squared is 0.628.

Linear regression on these six variables gives a fit with an R^2 value of 0.63 with error bars calculated by the QR method (regress function in Matlab [30]). The predicted against measured plot in Figure 7 shows a large amount of variance in the data that has not been captured by the linear model. The plot of residuals in Figure 8 shows some possible structure that may be due to non-linearities in the data.

A visualisation of the activity may be obtained by projecting the input parameters onto a 2-dimensional Kohonen self-organising feature map, an unsupervised form of Neural Network [31]. The co-ordinates on the Kohonen map are smoothed across the nodes by taking a mean inverse distance to the three nearest nodes:

$$(x, y) = \frac{w}{\text{sum}(w)} \cdot C,$$

where $w = [1/\text{d}1, 1/\text{d}2, 1/\text{d}3]$ are the inverse distances of a point to $C = [(x1, y1); (x2, y2); (x3, y3)]$, the three nearest nodes. The third dimension, or height above the map is set to be activity and a surface is obtained by triangle based linear interpolation (griddata function in Matlab [30]). The surface plotted in Figures 9 and 10 shows a degree of regularity, indicating that there may well be a (non-linear) relationship between molecular parameters and activity.

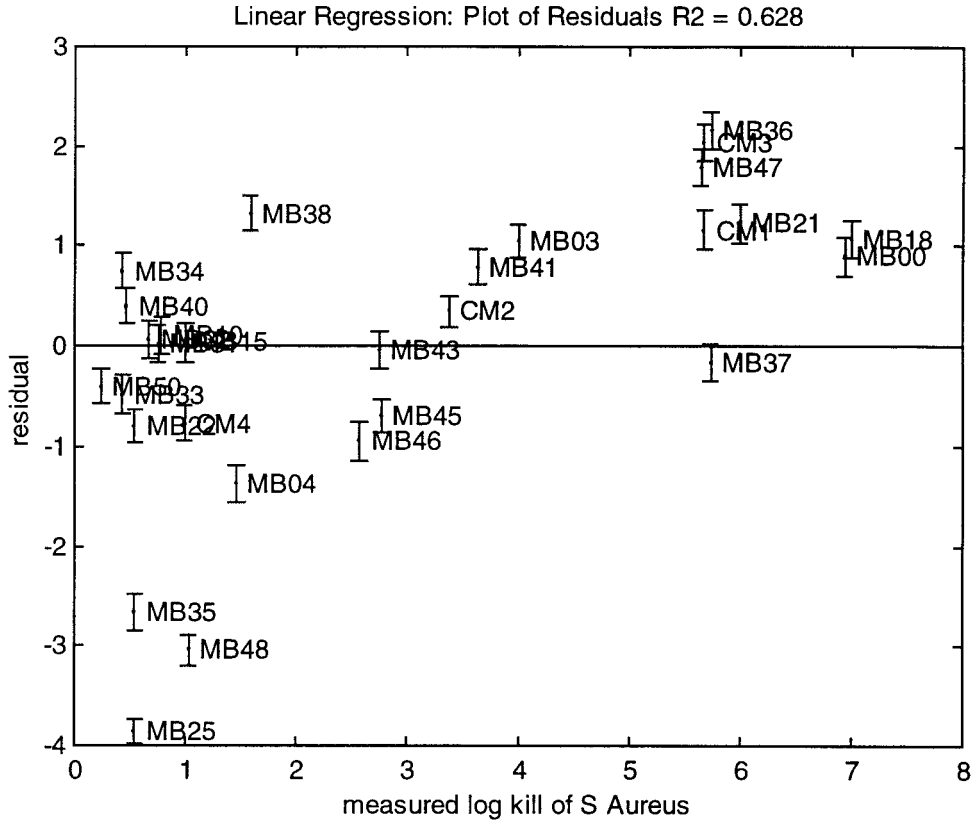

Figure 8. Plot of the residuals of the linear regression fit, showing the residual against the measured log kill of *S. aureus*. The trend in the residuals is indicative of non-linearity in the data.

Multi-Layer Perceptron (MLP) Neural Networks are described as universal non-linear approximators and have the ability to model arbitrary non-linearities [32], i.e. may be used to find non-linear correlations between inputs x to outputs y. For a univariate output y, this has a general form:

$$y = \sum_j (w_j \cdot (f(\sum_i w_{ij} x_i + b_i)) + b_j),$$

where: x_i are the inputs (general molecular and amphiphilic parameters); w_{ij} are the weights between the input layer and the 'hidden' layer; w_j are the weights between the 'hidden' layer and the single output node; b_i and b_j are the bias weights, or offsets; f is a non-linear transfer function, in this case the hyperbolic tangent tanh, though it may a sigmoid function or a Gaussian.

Classically, training an MLP consists of setting the weights w, such that the mean squared error $\Sigma(t_n - y_n)^2$ is minimised over n, an index to all the data points. This was done using the delta-bar-delta training algorithm in Neural Works Professional II [33]. A Neural Network with 6 hidden nodes was found to give the best results.

As the data set was very small (29 data points), a leave-one-out cross-validation technique was used to ensure the validity of the model. Gathering the predictions across all 29 points, it is possible to derive R^2 values for the cross-validation, as well as mean R^2 for the 29 models (Figure 11)

Mean R^2 over 29 models	0.968,
Cross-validated R^2 for *S. aureus*	0.905.

Recent results suggest that weighted combinations of models may outperform any single model [34, 35]. With no a priori reason to preferentially weight any one of the models, the predicted activity was taken to be the mean of the predictions of the 29 models.

Optimisation

Genetic Algorithms (GA) [15, 16] are currently of great interest as general-purpose optimisation tools.

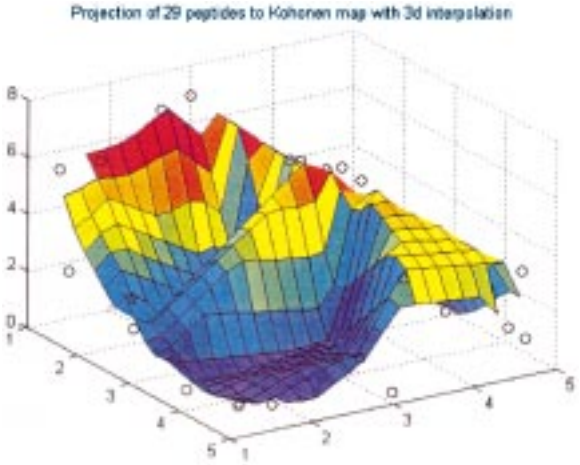

Figure 9. Projection of the 6 input molecular parameters for the 29 peptides projected onto a Kohonen map with 25 nodes. The projection is smoothed between the three nearest nodes. The height above the map is given by the measured log kill of *S. aureus*, and this is interpolated between the nodes to give an approximate 'landscape' of activity, equivalent to the contours in Figure 10. This enables an easy visualisation of the degree of regularity within an approximation to the true landscape.

Table 1. Efficiency comparison of the Genetic Algorithm against other standard optimisation techniques. The GA is an order of magnitude more efficient

Method	Efficiency
GA	$> 90/1250 = 7.2\%$
Monte Carlo	$1/200 = 0.5\%$
Random	$4 / 52000 = 0.008\%$

Table 2. Experimental measures of log kill of *S. Aureus* for peptide sequences that were generated by the Genetic Algorithm

Peptide sequence	Predicted log kill	Measured log kill
AASKAAKTLAKLLSSLLKLL	7.22	> 5.06
LLKKLLRAASKALSLL	7.13	> 5.90
AAKKLSKLLKTLLKLL	7.35	> 5.76
KALKKLLKLASSLLTAL	7.04	5.90
AASKALRTASRSLLTLL	7.03	> 5.85

Their use in peptide design is particularly apt due to the natural extension of biological evolutionary operators to amino acid representations of peptide sequences [4].

There are three main aspects to be determined in the devising of a Genetic Algorithm: the representation of the problem; the fitness function or 'chance of survival'; and the genetic operators, or method of mating. The problem is naturally represented by the sequence of amino acids, using single letter symbols: LLKALL…, etc. The fitness function, or method by which 'survival' of any individual is determined, is simply the Neural Net predictive model: 'virtual' peptides which are predicted to be good bactericides survive to produce offspring for later populations. As the peptides had to be effective against both bacteria, the fitness value assigned to a peptide was the lower of the *S. aureus* and *E. coli* predicted log Kill scores.

For this study standard operators: roulette wheel selection, mutation, and two-point crossover were used. Following a number of short tests on a simplified model, it was noted that the success of the algorithm in finding large numbers of potentially active peptides was robust to the setting of the GA parameters. For a full run, the parameters were set as follows: population size 100, size of elite 25, probability of cross-over 0.6, and probability of mutation 0.033.

Results

For an average GA run, after 50 generations, 90 out of 100 peptides in the final population were found to be acceptable, i.e. with predicted log kill >7 to both *S. aureus* and *E. coli*, given a total of $50 * 25 = 1250$ evaluations. This compares very favourably with random generation of peptides, 4 hits out of 52 000 evaluations, and Monte Carlo optimisation, 200 evaluations, averaged over 95 runs, to result in a single peptide. Note this efficiency is a minimum comparison: once the GA converged to a region of high activity, approximately 50% of new peptides generated by the algorithm were found to be active. Comparative results are given in Table 1.

Using the Genetic Algorithm, over 400 potentially active 'virtual' peptides were generated. Five of the 400 were selected to be synthesised by determining the Principal Components of the six chosen molecular descriptors of the 400, and selecting 5 peptides that maximised the diversity. The diversity in molecular properties is also reflected in the diversity of the amino acid sequences. The bactericidal properties of these five were measured and are given below:

The measured log kills given in Table 2 are given as lower bounds due to a threshold in the measurement sensitivity that is dependent on the initial cell numbers of *S. aureus* bacteria in the test innoculum.

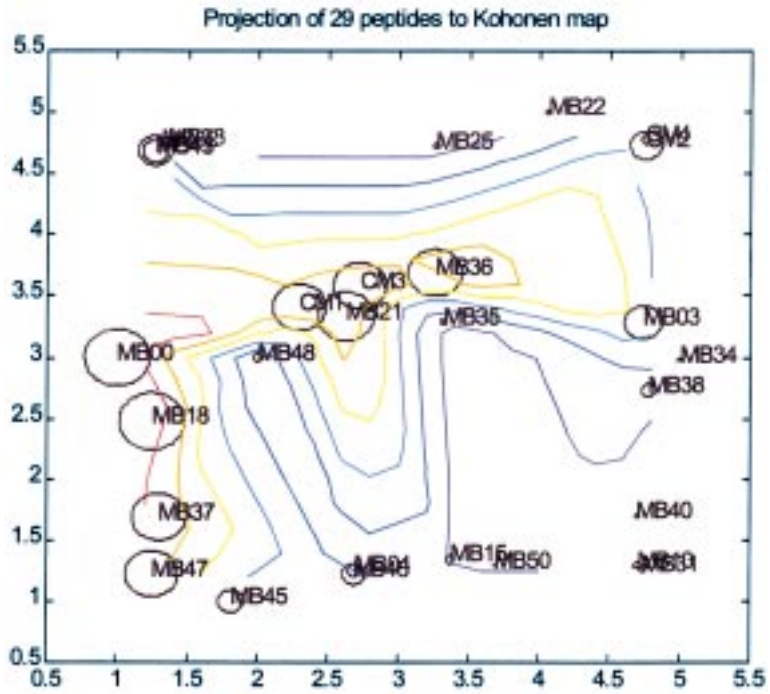

Figure 10. Projection of the 6 input molecular parameters for the 29 peptides projected onto a Kohonen map with 25 nodes. The projection is smoothed between the three nearest nodes. The radius of each circle represents the measured log kill of *S. aureus* of each peptide, and the contour lines interpolate that activity across the map. The regularity of the shape of the contours indicates that there is likely to be regularity in the pattern of bactericidal activity with regard to the molecular parameters.

Patent application

To be patentable an invention must be novel, useful and non-obvious. In addition, it should be understandable by a person, or group of people, 'skilled in the art'. Typically a patent application ends with a number of separate but related claims. The claim sets out what is covered by the patent, i.e. the particular intellectual property that is protected. There are a number of typical conventions in framing patents concerning peptides. A common method is to frame a claim as consisting of a list of individual peptides as given by their amino acid sequences, e.g. Ala-Leu-Thr-.... or by abbreviation LLLKLKKALL...., etc. A wider form of patent claim is based on the structural form of the peptide and defines templates in which amino acids may be substituted. A claim may read: peptides of the form 'R1-R2-R2-R1-R3- ... etc.' where the R-groups may be defined as hydrophilic amino acids, hydrophobic amino acids, basic hydrophobic, neutral hydrophilic, etc. [17, 36–38].

A typical example is found in Claim 1 of [37]:
'[we claim...] a biologically active amphiphilic peptide, said peptide including the following basic structure x, wherein x is:
R1-R1-R1-R3-R5-R1-R1-R1-R1-R1-R2-R2-R1-R1-R3-R1-R4-R 1-R3-R4-R1-R1
Wherein R1 is a hydrophobic amino acid, R2 is a basic hydrophilic amino acid, and R3 is a neutral hydrophilic amino acid, R4 is a hydrophobic or basic hydrophilic amino acid, and R5 is a hydrophobic, basic hydrophilic, or neutral hydrophilic amino acid'

The scientific justification for this claim is given by experiments conducted on 7 peptides:

GVLSNVIGYLKKLGTGALNAVL
GVLSKVIGYLKKLGTGALNAVL
GVLSQVIGYLKKLGTGALNAVL
GVLSFVIGYLKKLGTGHLNHVL
GVLSNVIGYLKKLGTGKLNKVL
GVLSFVIGYLKKLGTGKLNKVL
GVLSKVIGYLKKLGTGKLNKVL

where these peptides have been shown to be active bactericides, and an implicit inference is made that this bactericidal property will extend to other peptides with a similar hydrophobic – hydrophilic structure.

Let us compare this with [17]. The patent text states:
'We have determined that effective peptides are dis-

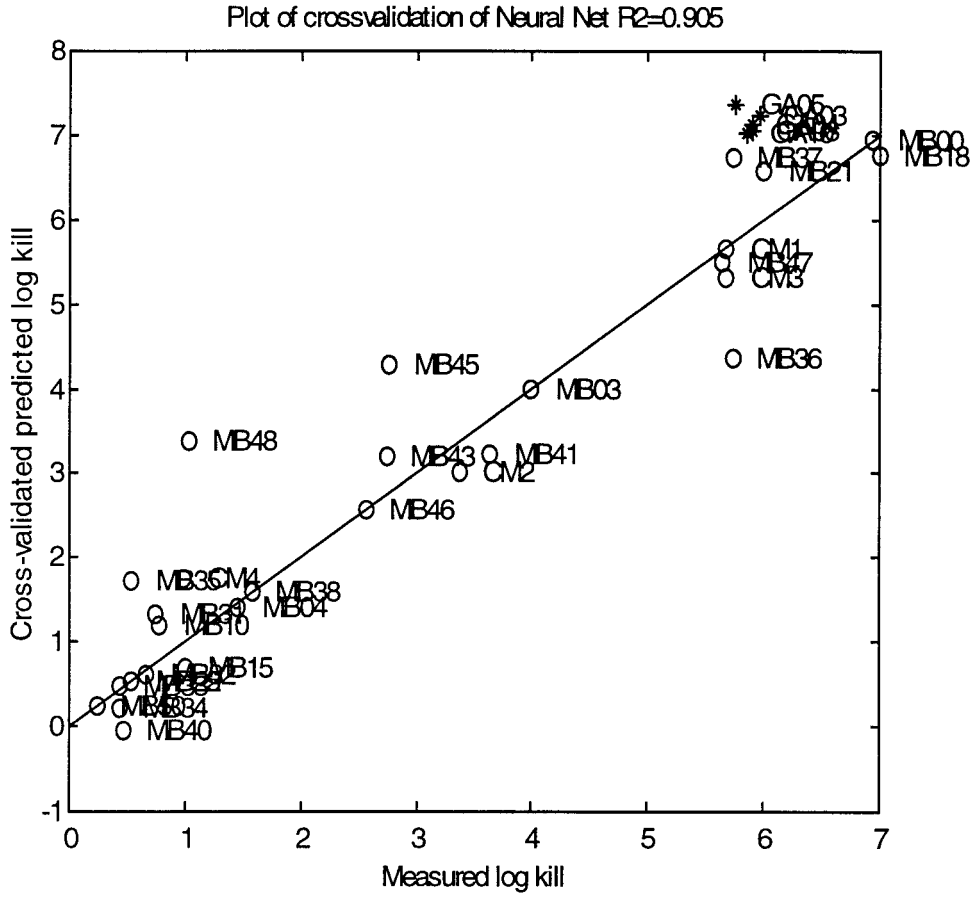

Figure 11. Results of the cross-validation of the Neural Network showing cross-validated prediction of log kill of *S. aureus* against measured value, circles (o) indicate the cross-validation set of 29 peptides, and the asterisks (∗) indicate the 5 'computer-designed' peptides. The results lie reasonably close to the line of identity with an *R*-squared of 0.905.

criminated from ineffective peptides by means of an equation relating certain properties of the peptides to their biological activity against specific micro-organisms... and have demonstrated that a strong correlation is shown between those peptides which satisfy this rule and those which have effective antimicrobial properties. Conversely, it appears that the majority of the peptides which exhibit properties outside the scope of this rule do not show effective antimicrobial properties.'

As publishing 29 Neural Net models would be somewhat cumbersome, for simplicity of publication in the patent, a single Neural Net model was generated that approximated the same function as the mean of the 29 Neural Nets. An artificial data set was generated by creating a lattice of $7^6 = 117,649$ data points with 'observed values' given by the mean of the 29 Neural Net models. This artificial data was used to train a single Neural Net model with $R^2 = 0.982$ over the artificial data set (Figure 14). This final model was published in the patent application: Claim 3 of [17] claims the peptides with specific bactericidal ability to *S. aureus*:

'Antimicrobial peptides having a length of 10–30 amino acid residues wherein the predicted log kill to S. aureus ATCC 6538, ($L_{S.\,aureus}$) is greater than 5, $L_{S.\,aureus}$ being given by the equation...'

and there follows several pages of simple algebraic expressions of the Neural Network, written in a form that may easily be turned into a computer programme.

While the Neural Network, and the necessary derivations of the input parameters to the Neural Network are disclosed in their entirety, the object of the patent is to protect the peptides that the NN predicts to have an activity higher than a specified value. In effect the patent states that we have determined a relationship

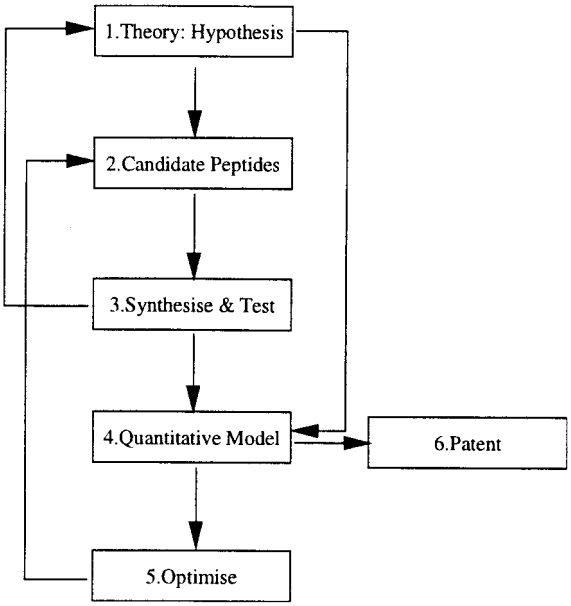

Figure 12. By using the Neural Net predictive model to define the scope of a patent claim, we have added a sixth step to the Computer Aided Molecular Design process, the interplay between modelling and patenting.

between peptide 3-D structure and bactericidal activity. This rule is given by the Neural Network model, and is validated by both positive and negative examples. The generalisation from specific examples to a wider range of peptides is explicit and is demonstrated to be based on a combination of scientific understanding and experimental evidence. Furthermore it is substantiated by the success in predicting new peptides, the sequences given in the results, that are indeed active against *S. aureus*. Many thousands of active peptides may easily be found using the Model Inversion / Optimisation techniques described here and elsewhere. It is important to note that the Neural Network is *not* patented or claimed in any way.

In effect we have added a sixth step to the process of Computer-Aided Molecular Design – the close interplay of the modelling step with the patenting process (Figure 12).

Comparing patent spaces

Whether a patent is written in any of the available forms [17, 36–38], a peptide patent claim may be viewed as defining a region in peptide space, this may be called the 'patent space'. It is instructive to analyse the shape and size of a patent space for a particular

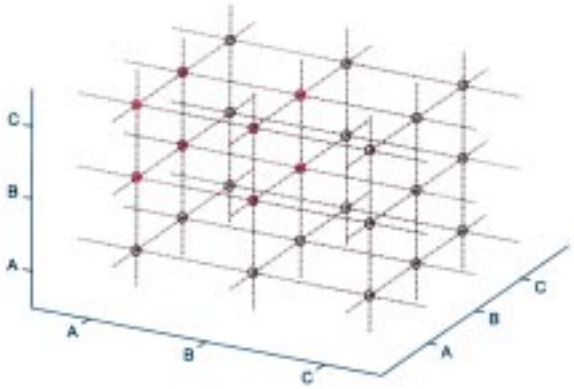

Figure 13. Schematic of a regular sub-space of a Peptide Space of trimers based on an amino acid alphabet of {A, B, C}. The sub-space is defined by R1-R1-R2 where R1 = {A,B}, and R2 = {B,C}.

claim or patent as well as the spread of illustrative examples through this space, and compare across patents. A small example may help to illustrate this point. For a 'toy' peptide system consisting of trimers with an amino acid alphabet of {A,B,C} the peptide space consists of a fully connected three-dimensional hyper-cube. If we take as a hypothetical case a peptide patent that claims R1-R1-R2 where R1 = {A, B} and R2 = {B, C}. The full set of sequences in this patent space is AAB, ABB, BAB, BBB, AAC, ABC, BAC, BBC, and they occupy a connected sub-cube of the complete space (Figure 13).

For [36–38] the shape of the patent space will be a fully connected hyper-cube nestling within the complete peptide space. The 'landscape' defined by a Neural Network model is more complex, and approaches to studying these are given in [7–9, 39, 40]. We wish to compare estimates of the relative extent of the experimental space (the diversity of the synthesised peptides) with the extent of the patent spaces claimed. A crude measure of the extent of a peptide (sub-) space is the average hamming distance between the peptides of the region, where the hamming distance between sequences S1 and S2 is the smallest number of amino acid substitutions or additions necessary to transform S1 into S2. The average hamming distance between the points of such a regular subspace may easily be calculated. For a given sequence N amino acids long, with A possible amino acids, the peptides that are s steps away involve $(A\text{-}1)^s$ changes, for which there are ${}_N C_s$ permutations. Summing from 0 to N, D the average hamming distance of the space is given by:

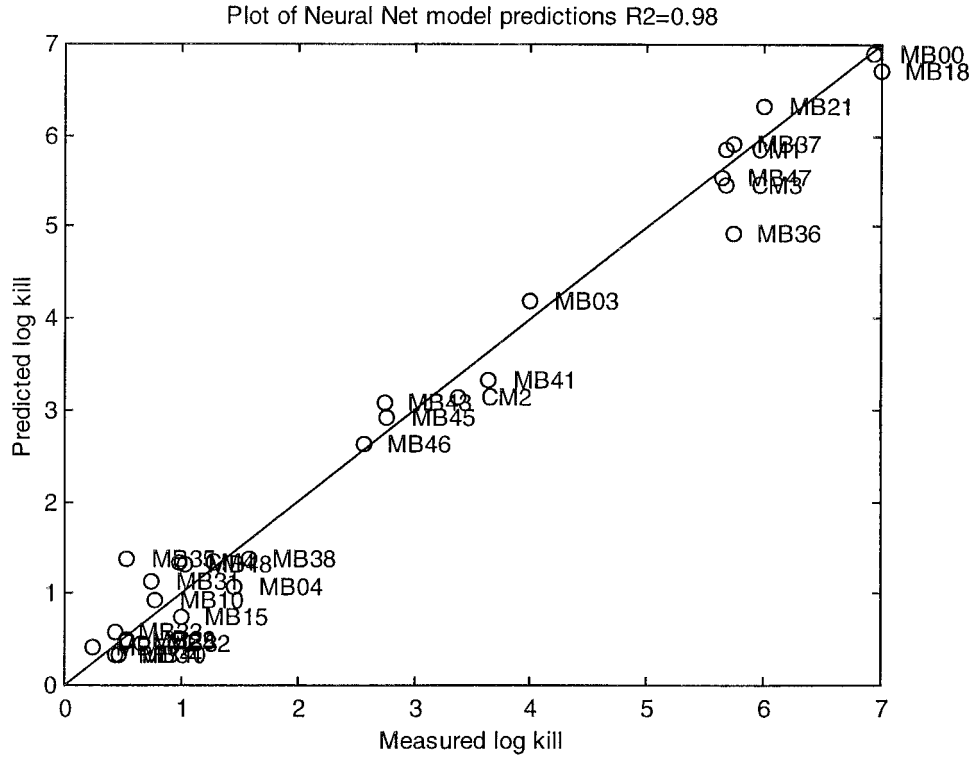

Figure 14. Plot of Neural Network model predictions based on the initial 29 synthesised peptides. The results lie very close to the line of identity with an *R*-squared of 0.98.

$$D = \frac{1}{T}\sum_{s=0}^{N} s.{}_{s}^{N}C(A-1)^{s},$$

where $T = A^N$ is the total number of peptides in the space. For [36–38], A was estimated at 10, half the typical alphabet of 20 amino acids. For [17] as there are no structural templates it was (over-)estimated at 20.

H, the average hamming distance between experimental examples was calculated directly from the examples for [17, 37, 38]. For [36], *H* was estimated by the substitution method of generating the example peptides described in the patent text.

Table 3 shows the comparison of the average hamming distance between the synthesised peptides given as examples and the average hamming distance of the patent space they fall within, where n = the number of example peptides within the patent space of the particular claim. The final column gives the comparison between the diversity of the synthesised peptides, and the diversity of the patent space claimed. This value gives a rough measure of the extent to which the experimental results extend across the patent space claimed, and may be used as justification for the claim. It can be seen that our patent [17] is at least comparable, if not more substantial, in its spread.

Table 3. The results of a comparison of the relative spread of synthesised peptides within the patent spaces claimed.

Patent	Claim	n	H	N	A	D	100∗H/D
[36] 5294605	7	41	2.03	27	10	16.2	12.5
[37] 92_20358	1	7	2.7	24	10	21.6	12.5
[38] 92_17197	1	15	9.45	26	10	24.3	38.9
[17] 96_28468	3	14	13.2	30	20	28.5	46.3

n = the number of example peptides within the patent space of the particular claim; H = the average hamming distance between experimental examples; A = an estimate of the number of possible amino acids in each position in the peptide sequence; D = the average hamming distance between peptides in the peptide patent space. The last column gives an indication of the relative spread within the claim.

Structurally based patents assume that peptide activity will be distributed in a highly regular manner through peptide space. In contrast a Neural Net based definition of patent space makes no prior assumptions about the shape of the claimed region. It is quite possi-

ble that there are disconnected regions or even isolated peptides. A more detailed study of the characteristics of these spaces is left to a further study.

Discussion and conclusions

Computer-Aided Molecular Design has taken a large leap forward in recent years by the use of algorithms such as Neural Networks for the prediction of properties of interest, and Genetic Algorithms that use the Neural Network models in the discovery and design of new peptides. This study shows how we have successfully used these techniques to design novel bactericidal peptides. Neural Net equations have been determined on the basis of a sound theoretical hypothesis of bactericidal activity, 29 experimental measures, carefully chosen molecular parameters and the development of a predictive model based on these aspects. Furthermore, this relationship has been validated by its use in computer-aided peptide design, i.e. new peptides have been designed by Genetic Algorithm that were predicted to be active, and synthesis and experimentation on these new peptides have been in accord with the predictions.

The patent aspects of the work have identified several legal issues. It appears reasonable to assume that where new and useful compounds are produced by a method involving intensive computation, those who set up the problem, rather than the computer itself, are the 'inventors' of the compounds, despite the fact that the precise nature of the compounds could not have been predicted in advance. We describe how we have used the Neural Net equations to define the scope of a patent claim. A justification for this approach lies within the context of the well-documented method of Computer Aided Molecular Design. It remains to be seen whether such a mathematical approach will rest easy with the Patent Offices around the world who must approve the patent for grant.

In the same way that any patent discloses the scientific understanding of a particular scientific endeavour, so we have disclosed the model in full, as well as the precise method for development of the model, so that it may contribute to the body of scientific knowledge in this domain. The mental act of formulating the claims of a patent, which identifies its scope, has always involved the definition of a broad space within which the examples and many other potential compounds lie. We have begun to show how the diversity of this space can be explored with mathematical tools. In a simple comparison with other patents on bactericidal peptides, we show that in our case the diversity of the synthesised peptides is extremely high. The residues vary in every position, and this paper presents some evidence to show that they have a relatively wide range within the space of all peptides. While these results are very interesting, considerable further research may be carried out on molecular patent space analysis to address more fully issues surrounding 'coverage' of example molecules and claims.

References

1. Venkatasubramaniam, V., Chan, K. and Caruthers, J. M., Comput. Chem. Eng., 18 (1994) 833.
2. Maranas, C.D., AIChE J., 43 (1997) 1250.
3. de Weijer, A.P., Lucasius, C.B., Buydens, L. and Kateman, G., Chemometrics Intelligent Lab. Systems, 20 (1993) 45.
4. Jones, D. T., Protein Sci., 3 (1994) 567.
5. Hellinga, H.W. and Richards, F.M., Proc. Natl. Acad. Sci. USA, 91 (1994) 5803.
6. Jin, A.Y., Leung, F. Y. and Weaver, D. F., J. Comput. Chem., 18 (1997) 1971.
7. Forst, C.V., Reidys, C. and Weber, J., In Moran, F. (Ed.), Advances in Artificial Life, Vol. 929, Springer-Verlag, Berlin, 1995, pp. 3628–4147.
8. Stadler, P.F., Santa Fe Institute paper 97-11-082, Santa Fe Institute, NM, USA.
9. Fontana, W., Stadler, P.F., Tarazona, P., Weinberger, E.D. and Schuster P., Phys. Rev. E, 47 (1993) 2083.
10. Schneider, G., Schuchhardt. J. and Wrede, P., Biol. Cybernetics, 73 (1995) 3.
11. Clark, D.E. and Westhead, D.R., J. Comput.-Aided Mol. Design, 10 (1996) 337.
12. Mavrovounitis, M.L., In Davis, J.F., Stephanopopoulos, G. and Venkatasubramanian, V. (Eds), Proc. Int. Conf. on Intelligent Systems in Process Engineering, Vol. 92, AIChE Symp. Series no. 312, 1996, p. 133.
13. Parrill, A., Exp. Opin. Ther. Patents, 7 (1997) 937.
14. Venkatasubramaniam, V., Sundaram, A., Chan, K. and Caruthers, J. M., Abs Papers of the Am. Chemical Soc., 207 (1994) 60-COMP part 1, 396.
15. Holland, J.H., Adaptation in Natural and Artificial Systems, Michigan University Press, Ann Arbor, MI, USA, 1975.
16. Goldberg, D.E., Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, USA, 1989.
17. Bhakoo, M., Patel, S. and Stott, I., Amphiphilic Peptides and Analogs Thereof, WO 96/28468 (1996).
18. Wade, D., Andreu, D., Mitchell, S.A., Silveira, A.M.V., Bowman, A., Bowman, H.G. and Merrifield, R.B., Int. J. Pept. Protein Res., 40 (1992) 429.
19. Christensen, B., Proc. Natl. Acad. Sci. USA, 85 (1988) 5072.
20. Hancock, I. C. and Poxton, I. R. (Eds), Bactericidal Cell Surface Techniques, John Wiley & Sons, Chichester, UK, 1988.
21. Veld, G.I., Drissen, A. J. M. and Konings, W.N., FEMS Microbiol. Rev., 12 (1993) 293.
22. Eisenberg, D., Weiss, R. M. and Terwilliger, T.C., Nature, 299 (1982) 371.

23. Eisenberg, D., Weiss, R. M. and Terwilliger, T.C., Proc. Natl. Acad. Sci. USA, 81 (1984) 140.
24. Degrado, W.F., Adv. Prot. Chem., 39 (1988) 51.
25. Bhakoo, M., Birkbeck, T.H. and Freer, J.H., Biochemistry, 21 (1982) 6879.
26. Bhakoo, M., Birkbeck, T.H. and Freer, J.H., Can. J. Biochem. Cell Biol., 63 (1985) 1.
27. Fasman, G. (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, 1989, Plenum Press, New York, NY.
28. Freer, J.H., Birkbeck, T.H. and Bhakoo, M., in Alouf, J.E. (Ed.), Bacterial Protein Toxins, Academic Press, London 1984, pp. 181–189.
29. Sybyl^{R}, TRIPOS Inc., St. Louis, MO.
30. Matlab^{TM}, The Mathworks Inc., Natick, MA.
31. Kohonen, T., Self-Organising Maps, Springer-Verlag, Berlin (2nd edn) 1995.
32. Hornick, K., Stinchcombe, M. and White, H., Neural Networks, 2 (1989) 359.
33. Neural Works Professional II^{TM}, NeuralWare Inc, Pittsburgh, PA.
34. Bishop, C., Neural Networks for Pattern Recognition, Clarendon Press, Oxford, UK, 1995.
35. Perrone, M.P. and Cooper, L.N., in Mammone, R.J. (Ed.), Artificial Neural Networks for Speech and Vision, Chapman & Hall, UK, 1993, pp. 126–142.
36. Maloy, W.L., Novel Peptide Compositions and Uses Thereof WO 92/17197 (1992).
37. Zasloff, M.A., Novel Biologically Active Peptide Compositions and Uses Thereof, WO 92/20358 (1992).
38. Houghten, R.A. and Blondelle, S., Amphiphilic Peptide Compositions and Analogs Thereof, US Patent 5,294,605, 1994 (filed: July 8 1991).
39. Kauffman, S.A. and Macready, W.G., J. Theor. Biol., 173 (1995) 427.
40. Levitan, B., Annu. Rep. Comb. Chem. Mol. Div., 1 (1997) 95.

Appendix 1. Primary peptide sequences with measured and predicted activity

Example	Sequence	*S. aureus*		*E. coli*	
		Act	Pre	Act	Pre
MB_03	VSSKYLSKVKVKAGK	4	4.18	1.63	0.75
MB_04	ARLAKKALRRLAKKD	1.46	1.08	0.83	0.44
MB_10	GESLASKAAKAAER	0.78	0.92	0.46	0.45
MB_15	ESLAKALSKEALKALK	1	0.74	1.24	0.44
MB_18	**LKALKKLAKKLKKLA**	**7**	**6.70**	**4.17**	**6.28**
MB_22	GWLLLEYIPVIAAL	0.54	0.49	0.41	0.44
MB_31	EAALKAALDLAAKLA	0.75	1.13	0.39	0.45
MB_00	LKLLKKLLKKLKKLL	**6.94**	**6.90**	**7.15**	**6.02**
MB_21	FASLLGKALKALAKQ	**6**	**6.32**	**6.15**	**5.68**
MB_25	LSSALSALSSALSSK	0.54	0.46	0.37	0.44
MB_32	ERSAAKSAARSLARR	0.67	0.46	0.08	0.44
MB_33	EKTLARTAAKTALKK	0.43	0.58	0.22	0.44
MB_34	EKAAAKSAAAKTLARR	0.43	0.33	0.24	0.44
MB_35	VSSKYLSKALVKAGR	0.54	1.38	0.26	0.45
MB-36	**FASLLGKALKALLAKLAKQ**	**5.74**	**4.92**	**5.95**	**5.40**
MB-37	**FASLLGKLAKKLAKKALK**	**5.74**	**5.91**	**5.22**	**4.96**
MB-38	ESLKARSLKKSLKLKKLL	1.58	1.38	0.58	0.45
MB-41	ELAKKALKALKKALKSAR	3.64	3.33	0.18	0.59
MB-43	ELAKKALRALKKALKSAK	2.75	3.07	0.22	0.54
MB-45	ETFAKKALKALEKLLKKG	2.77	2.91	0.16	0.57
CM-1	**LALLKVLLRKIKKAL**	**5.68**	**5.85**	**4**	**3.54**
CM-2	LULLLKILLLKKLKA	3.38	3.14	0.75	0.83
CM-3	**ALKAALLAILKIVRVIKK**	**5.68**	**5.47**	**3.07**	**4.07**
CM-4	LLAILLLALLALRKKVLA	0.99	1.34	0.38	0.46
MB-40	ETELAKKALKALKLKKLA	0.47	0.32	0.16	0.44
MB-46	ESSLKKKALSKLSKLLKKG	2.57	2.63	0.28	0.51
MB-47	**QKAASRLLRALSKLLEAF**	**5.65**	**5.55**	0.11	0.48
MB-48	QKALAKLAKKALKALAKQ	1.03	1.31	1.77	0.45
MB-50	ESKAAKAAKKAAKAKASE	0.24	0.40	0.18	0.44
MC-03	**AASKAAKTLAKLLSSLLKL**	**5.96**	**7.24**	1.96	1.72
MC-04	**LLKKLLRAASKALSLL**	**5.9**	**7.13**	0.45	0.82
MC-05	**AAKKLSKLLKTLLKLL**	**5.76**	**7.36**	1.01	2.06
MC-08	**KALKKLLKLASSLLTAL**	**5.9**	**7.06**	1.61	1.12
MC-10	**AASKALRTASRLARSLLT**	**5.85**	**7.03**	0.41	−0.56

Appendix 2. The 39 molecular parameters calculated for each peptide

Molecular parameter class	Number of parameters	Molecular parameters
Simple	4	Molecular weight, length of the alpha helix, cross-sectional area of the alpha helix and length/weight
Hydrophobic dipole	5	*x*, *y*, *z* and *zy* components and magnitude
Hydrophobicity	3	Sum of the hydrophobicity of the hydrophobic amino acids, sum of the hydrophobicity of the hydrophilic amino acids and the sum of these two parts
Solvent accessible surface area	3	Hydrophobic, hydrophilic and hydrophilic/ hydrophobic
Charge dipole	4	*x*, *y* and *z* components and magnitude
Charge	4	Sum of the charge of positively charged amino acids, the sum of the charge of positively charged amino acids, the sum and the difference of these two parts
Fit of hydrophobicity of amino acids to different patterns	2	Fit to sine wave and fit to square wave
Radius of gyration	2	*Rxyz* and *Rxy*
Radial moment	2	The radial moment and the square of the radial moment
Number of atoms between r and $(r + 1)$ Angstroms from the alpha helix axis	10	For $r = 0$ to 9