BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 11 2005, pages 2651–2656
doi:10.1093/bioinformatics/bti405

Sequence analysis

Genetic algorithm-based optimization of hydrophobicity tables

Moti Zviling, Hadas Leonov and Isaiah T. Arkin*

The Alexander Silberman Institute of Life Sciences, Department of Biological Chemistry, The Hebrew University, Givat-Ram, Jerusalem 91904, Israel

Received on August 1, 2004; revised and accepted on March 22, 2005
Advance Access publication March 29, 2005

ABSTRACT

Summary: The genomic abundance and pharmacological importance of membrane proteins have fueled efforts to identify them based solely on sequence information. Previous methods based on the physicochemical principle of a sliding window of hydrophobicity (hydropathy analysis) have been replaced by approaches based on hidden Markov models or neural networks which prevail due to their probabilistic orientation. In the current study, an optimization of the hydrophobicity tables used in hydropathy analysis is performed using a genetic algorithm. As such, the approach can be viewed as a synthesis between the physicochemically and statistically based methods. The resulting hydrophobicity tables lead to significant improvement in the prediction accuracy of hydropathy analysis. Furthermore, since hydropathy analysis is less dependent on the basis set of membrane proteins is used to hone the statistically based methods, as well as being faster, it may be valuable in the analysis of new genomes. Finally, the values obtained for each of the amino acids in the new hydrophobicity tables are discussed.

Availability:

Contact: arkin@cc.huji.ac.il

1 INTRODUCTION

Membrane proteins are of major importance and interest as they participate in the transduction of signals across the membrane and control the flow of molecules in and out of the cell. Their importance is reflected by their genomic abundance, and in a wide variety of organisms, membrane proteins constitute between 20 to 35% of all proteins in their genome (Wallin and von Heijne, 1998; Stevens and Arkin, 2000b). Moreover, they are extremely important in biomedicine as they serve as targets for most pharmaceutical agents in clinical use today.

Structures solved so far, permit classification of membrane proteins into two categories: α-helical bundles and β-barrels (White *et al.*, 2001). Since α-helical bundles are far more prevalent in the genome (Wallin and von Heijne, 1998; Stevens and Arkin, 2000b), and their biomedical importance is greater than that of β-barrels, all discussion henceforth will be concerned only with membrane proteins which adopt an α-helical bundle fold.

The importance of membrane proteins has promoted efforts to develop algorithms capable of accurately predicting their presence based on sequence information. Programs in common use reach appreciable prediction levels (Kall and Sonnhammer, 2002), and include TMHMM (Sonnhammer *et al.*, 1998), HMMTOP (Tusnady and Simon, 2001), PHDhtm (Rost *et al.*, 1996) and MEMSAT (Jones *et al.*, 1994). These programs are based on statistical characteristics of transmembrane helices. For example, TMHMM and HMMTOP are based on hidden Markov models which provide a probabilistic framework for many different types of problems. This framework consists of a set of states corresponding to the general biological scheme of the protein regions that are modeled. For instance, models of membrane proteins will generally include states for cytoplasmic loops, transmembrane regions and periplasmic loops. Each state is characterized by the distribution of amino acids in the region, its models and by the state-transition probabilities. Hidden Markov models require a training period in which the parameters of the model are estimated with various algorithms, and then, assuming it was successful, the model can be used to classify any given input.

A different approach is used in the program PHDhtm. When given a protein sequence, it searches for sequence homologues and uses neural networks to predict transmembrane regions from the multiple sequence alignment.

Prior to the development of statistically based methods, hydropathy analysis pioneered by Kyte and Doolittle (1982), had been used to detect transmembrane regions. It relied on the fact that transmembrane α-helices are generally characterized by a hydrophobic stretch of $\sim$20 amino acids. The method is implemented by sliding a window of a given size along the sequence (e.g. 20 amino acids) and summing the hydrophobicity values of the amino acids within the window. If the sum of the values for any window is above a certain threshold, it is recorded as a membrane spanning region.

Further improvements to hydropathy analysis were introduced by von Heijne and co-workers in the program TopPred (von Heijne, 1989; Claros and von Heijne, 1994). This method combines hydropathy analysis of helices with a search for positively charged amino acids in the cytoplasmic juxtamembranous regions of transmembrane α-helices.

The necessity of determining the values that will predict transmembrane regions most efficiently had given rise to several hydrophobicity scales. Kyte and Doolittle (1982) used both the water-vapor transfer free energies and the interior–exterior distribution of amino-acids (Chothia, 1976). Engelman *et al.* (1986) determined the free energy for each of the 20 amino acids when transfered from water to oil by calculating the surface area of each amino acid side chain in an α-helix. An experimental scale was obtained by White and co-workers based on the transfer free energies for each amino acid in a pentapeptide from *n*-octanol to water (Wimley *et al.*, 1996).

Hydropathy analysis, unlike statistical methods, does not rely upon a training set of known membrane proteins for calibration. In other words, the physical criteria that determine the stability of

*To whom correspondence should be addressed.

membrane proteins are expected to be relatively constant among different organisms. In contrast, statistical sequence characteristics that may be unique to a specific organism, such as nucleotide bias (Stevens and Arkin, 2000a), may lead to poorer prediction in whole genome analyses if the training set was taken from different organisms (Kall and Sonnhammer, 2002).

In the current study, we present a statistical method that we have developed, which improves the prediction accuracy of hydropathy analysis. The method can be viewed as a synthesis between the statistically and physicochemically based methods. Using a genetic algorithm (GA), hydrophobicity tables with improved prediction accuracy were selected, based on a large dataset of structurally determined transmembrane proteins as well as aqueous proteins. The Matthew's correlation coefficient was used as a rigorous statistical marker, employing cross-validation, in order to show that the predictive power of the resulting hydrophobicity table improved appreciably. We also compare the hydrophobicity values obtained with those of previous methods and provide insight on the contribution of individual amino acids to more accurate predictions of transmembrane helices.

2 METHODS

We have used a GA incorporating different window sizes and thresholds in order to optimize known hydrophobicity tables over a set of solved membrane and aqueous proteins, thereby improving the predictive power of the hydropathy analysis.

2.1 Construction of the datasets

When constructing training and test sets, it is important to select proteins with unambiguous topology assignments. Since such an assignment is possible only when selecting proteins whose structure has been determined, all non-homologous α-helical membrane proteins in the list given by White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) were downloaded from the protein data bank (Berman *et al.*, 2000). Furthermore, since accurate prediction involves minimization of false positives (water-soluble proteins predicted to be membrane proteins), a database of water-soluble proteins was constructed from a non-redundant set of the protein data bank (Holm and Sander, 1996, 1997). The ratio of membrane proteins to water-soluble proteins in our training and test sets reflected the genomic ratio of the above, which is roughly 1:3 (Stevens and Arkin, 2000b).

The training set contained 85 membrane protein chains and 255 water-soluble protein chains. The transmembrane segments within the set of membrane proteins were defined according to the original structure publication, by analysis with DSSP (Kabsch and Sander, 1983) and by visual inspection of the structure. The test set, used to measure the performance of the algorithm, was built in a similar manner. It consisted of 8 membrane protein chains and 24 water-soluble protein chains, comprising ~10% of all protein chains used in this study.

2.2 Hydropathy analysis

As stated above, hydropathy analysis involves summation of the hydrophobicity values which are represented by free energy values. Hence, each window summation results in a ΔG value that states its overall hydrophobicity. The results are presented as a graph of sequence positions versus $\Delta G_{\text{Water}\rightarrow\text{Oil}}$. Each point represents the hydrophobicity of a putative transmembrane helix starting at the given position along the sequence (example in Fig. 1). If a peak in the graph exceeds a certain empirically chosen threshold and does not overlap with any other transmembrane segment, it is viewed as representing the start of a new putative transmembrane α-helix. In the case where an overlap occurs, both helices are combined into one larger putative α-helix. An overlap between helices is considered when the difference between their positions is less than a quarter of the window size used.

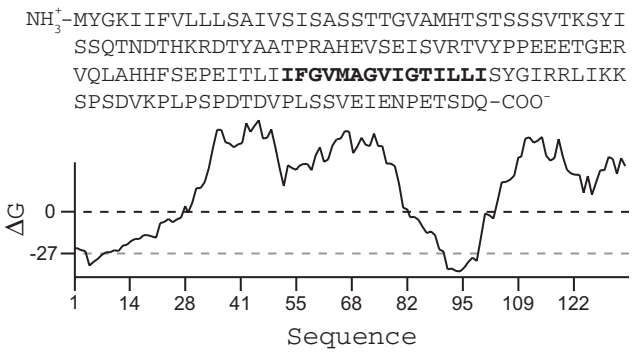

Fig. 1. Hydropathy analysis of human glycophorin A, the first membrane protein to be sequenced (Tomita and Marchesi, 1975). Calculation was done according to the GES scale hydrophobicity table (Engelman *et al.*, 1986), with a threshold of $\Delta G_{\text{Water}\rightarrow\text{Oil}} = -27$ kcal/mol (shown as a gray dotted line) and a window size of 15 amino acids (Stevens and Arkin, 2000b). Bold letters indicate the region which reaches maximum hydrophobicity according to the above criteria.

2.3 Scanning the training set

The program we developed accepts as input a hydrophobicity table ($H\Phi_i$), a hydrophobicity cutoff, a sensitivity parameter (see below) and a window size. Subsequently, the program performs a hydrophobicity analysis on each of the sequences within the training set (membrane and soluble proteins) by sliding the window over each sequence.

2.4 Score calculation

In any prediction process there can be four different possibilities to account for:

TP: True positive, a transmembrane helix correctly predicted.

FP: False positive, a transmembrane helix predicted when there was none.

TN: True negative, no transmembrane helix predicted when there was none to predict.

FN: False negative, no transmembrane helix predicted when there was one to predict.

It is clear therefore, that any single number that represents the predictive power of the method must account for all of the possibilities listed above. One such factor is the Matthew's correlation coefficient, given by

$$C \equiv \frac{TP\,TN - FN\,FP}{\sqrt{(TN+FN)(TP+FN)(TN+FP)(TP+FP)}} \qquad (1)$$

The Matthew's correlation coefficient ranges from $-1 \leq C \leq 1$. A value of $C = 1$ indicates the best possible prediction, in that every transmembrane helix was correctly predicted, and only true transmembrane helices were predicted. A Matthew's correlation coefficient of $C = -1$ indicates the worst possible prediction (or anti-correlation), where not a single transmembrane helix was correctly predicted and the largest number of incorrect transmembrane helices were predicted. Finally, a Matthew's correlation coefficient of $C = 0$ would be expected for a random prediction scheme.

Every predicted helix from the training set was classified as true/false positive according to a sensitivity parameter that ranged from 0 to 6. *TP* was defined when the start of a predicted transmembrane helix was within ± 0 to 6 amino acids from the start of the actual one, and *FP* was defined otherwise.

TN was defined as the potential number of non-transmembrane segments in the protein. In a water-soluble protein it was equal to the size of the protein divided by the window size and truncated to the nearest integer, where as in a membrane protein, every extramembraneous region was considered as an independent water-soluble protein.

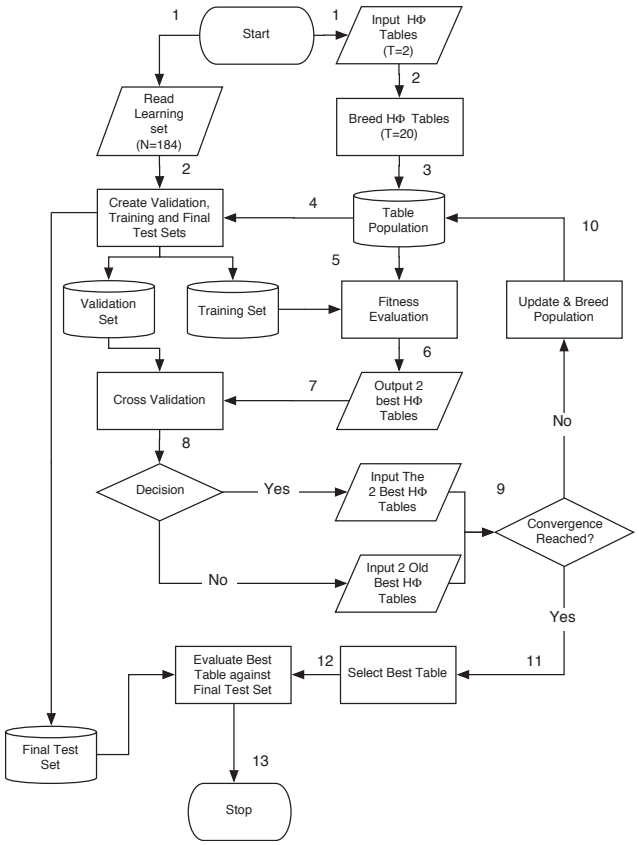

Fig. 2. Schematic diagram of the GA used to enhance the predictive value of hydrophobicity tables. See text for details.

The Matthew's correlation coefficient was calculated for each sequence in the training set, and the values were averaged to get an overall value for the entire set. The process was repeated for a combination of window sizes from 15 to 25 and hydrophobicity cutoff values of 20 to 35.

2.5 Genetic algorithm (GA) scheme

Genetic algorithms are adaptive heuristic search algorithms (Holland, 1975). As such they represent an intelligent exploitation of a stochastic search within a solution population which is used to solve optimization problems. They also assign a fitness value to each member of the population and behave in an analogous manner to Darwinian evolution: they strive to increase (i.e. optimize) the fitness of the individuals within the population (Koza, 1992). A GA was chosen over other optimization methods owing to the following reasons: (1) it allows the incorporation of objective information rather than derived or auxiliary knowledge, (2) it is considered an excellent search method where the solution search space is extremely large, and, most importantly, (3) the usage of randomized parameters helps to avoid local optimum solutions and contributes to the robustness of the algorithm.

2.5.1 GA description The GA we applied is based on the common roulette wheel selection technique (Fig. 2). The algorithm generates a population comprised of 20 hydrophobicity tables. It simulates the course of evolution which is expressed by breeding steps, in order to improve the fitness of individual tables within the population. The fitness is expressed by the number of correctly predicted membrane regions in a training set of proteins.

The algorithm adopts the following procedure. It starts with two hydrophobicity tables: Kyte–Doolittle scale (KD) (Kyte and Doolittle, 1982) and Goldman–Engelman–Steitz scale (GES) (Engelman *et al.*, 1986). This is shown in Figure 2, stage 1. The tables are used as a source for the initiation of the first population of hydrophobicity tables. Twenty random tables are generated from the initial tables (Fig. 2, stages 2 and 3) in a breeding process described in detail below. All generated tables are used to predict transmembrane regions in the training set, and are then evaluated for fitness (Fig. 2, stages 5 and 6). The two best tables (highest fitness value) are chosen for a cross-validation process (Fig. 2, stage 7). Depending on the results of the cross-validation process (see details below and in Fig. 2, stages 8 and 9) either the two current or the two previous tables (from the previous iteration) are chosen. At this point, if the algorithm has converged, the process is stopped and the final 200 best tables are tested on a final unseen test set (Fig. 2, stages 11–13). Otherwise, the selected tables are used in another iteration of the GA (Fig. 2, stages 9 and 10).

It is important to note that at each iteration of the algorithm new cross-validation and training sets are generated from a constant initial learning set of 340 proteins (Fig. 2, stage 4) and that the 200 best tables are selected for testing in order not to create a bias in the Mathew's correlation coefficient results subjected to the best table only.

2.5.2 Table encoding A 1D vector of size 20 was used for each table's encoding. Each cell in the vector represented a hydrophobicity value for a different amino acid, using a floating point representation, since it is preferable to represent the genes in a continuous, rather than in a discrete distribution. The genotype for the solution with k genes was symbolized as a vector $(x_1 \cdots x_k)$ with $x_i \in R$, where in this case, $k = 20$. The order of each amino acid's representation in the vector is shown in Table 3, where the first amino acid is Ile and the last one is Arg. In the KD table for instance, the hydrophobicity value for Ile is 4.5 and appears in the first cell, while the value for Arg is -4.5 and appears in the last cell.

2.5.3 Window-size and threshold The window-size and threshold used in the hydropathy analysis were evolved in a similar fashion to that of the values for the hydrophobicity of each amino acid. Specifically, every table contained two additional elements (21 and 22, respectively) that held the values for the window-size and hydrophobicity threshold. These elements were allowed to mutate (in changes of ± 1) and cross-over between tables in a similar fashion to all other table elements.

2.5.4 Generating the population Using a pair of hydrophobicity tables as parents, 20 siblings were generated in a breeding process that was repeated 10 times. Each included recombination and mutation steps, and produced two new hybrid tables (Fig. 3). The breeding process occurred at each iteration of the GA, and was carried out in the following manner:

- Cross-over stage (Fig. 3A–C): An even integer $2 \leq z \leq 20$ was randomly chosen and the number of exchange positions between the hydrophobicity tables was determined. $z/2$ represents the number of cross-overs. Then, a vector of exchange positions $(x_1 \cdots x_z)$, where $1 \leq x_i \leq 20$, was generated randomly for each of the two original tables. Finally, two new hybridized tables were produced by melding the information from both original tables as a simple recombination process.
- Mutation stage (Fig. 3D): two integers $1 \leq m$ and $n \leq 20$ were randomly chosen and the number of mutations to be performed on each of the tables was determined. Then, two vectors of mutation positions $(y_1 \cdots y_m$ and $w_1 \cdots w_n$, where $1 \leq y_i$ and $w_i \leq 20)$ were generated randomly. The table values that were chosen for mutation were changed by ± 0.5 according to a binary random decision.

2.5.5 Training and cross-validation In each iteration of the GA, the learning set was divided into two smaller sets: (1) a training set which consisted of 90% of the proteins from the learning set and (2) a validation set, which consisted of the remaining 10% of the learning set. The division was done by random choice. The training set was then scanned by hydropathy analysis, and the predictive power of the current population expressed by the Matthew's correlation coefficient was calculated for each

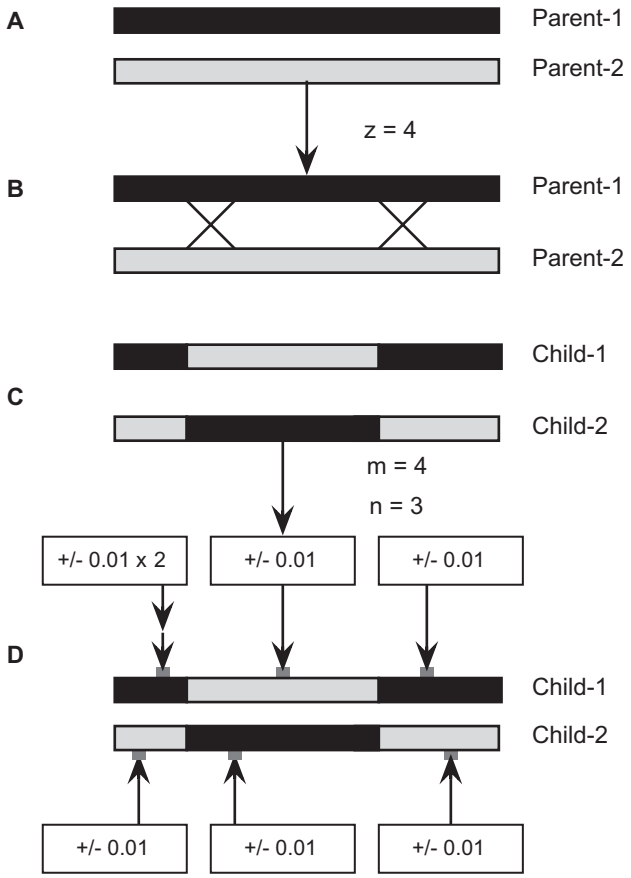

Fig. 3. Schematic diagram of the population generation process. (**A**) Two tables for further breeding are chosen, denoted by parent-1 (black shape) and parent-2 (gray shape). (**B**) A number of cross-over positions at each table is randomly chosen (e.g. $z = 4$), along with a random choice of the exchange positions themselves. (**C**) The new tables are shown after performing recombination. Each of the new tables is a product of hybridization between the original tables. (**D**) The number of mutations to be performed for each table is randomly chosen (e.g. $m = 4$ and $n = 3$), along with the mutation positions, resulting in the change of each chosen table value by ± 0.05. The same mutation position can be chosen more than once.

table. (Fig. 2, stage 5). Next, a cross-validation process that measured the progress of the training phase was performed by selecting the two best-predicting tables from the training phase and testing them on the validation set (Fig. 2, stages 6–8). Cross-validation was considered successful if the Matthew's correlation coefficient obtained from testing the two tables on the validation set was higher than the value of the previous round. In that case, if the algorithm did not converge or reach its maximum generation value (i.e. it was still possible to optimize the hydrophobicity values), another breeding process began (Fig. 2, stage 10), using these two tables as parents, and generating 20 siblings as described in Section 2.5.4. The old population of tables was replaced by the new siblings according to a replacement (selection) parameter.

The replacement parameter was determined in the program's initiation phase, and ranged from 20 to 80%. It determined what percentage of unfit tables remained in the training set, and how many of them were replaced. For example, a replacement of 20% represented a state in which 20% of the unfit tables had been replaced by new randomized tables. Tables were considered unfit if their Matthew's correlation coefficient was <0.5. If cross-validation failed, the two best optimized tables from the previous round were selected as parents for the next iteration's breeding process. Upon reaching convergence of the hydrophobicity tables or the maximum generation value, the process was stopped and the best 200 tables were chosen for the final evaluation. Finally, all parameters used in the algorithm are summarized in Table 1.

Table 1. GA and program parameters

Parameter	Value	Meaning
Table population	20	Number of tables to be used in the training phase
Population (input proteins)	340	Number of sequences in initial learning population
Training set	90%	Percentage of sequences from the learning population used during training
Validation set	10%	Percentage of sequences from the learning population used during cross-validation
Final test set	32	Number of sequences in the test population
Mutation units	±0.5	A point mutation change
Unfit tables threshold	0.5	The threshold defining unfit tables as replacement candidates
Selection parameter	20–80%	Replacement percentage of unfit tables
C value	−1 to 1	Matthew's correlation coefficient, an indication of a table's predictive power

Table 2. Prediction values measured on three different sets (training and test), according to the Matthew's correlation coefficient of hydropathy analysis and statistically based methods against GA

Prediction method	Training set			Test set		
	1	2	3	1	2	3
Hydropathy: GES	0.62	0.71	0.71	0.59	0.71	0.70
Hydropathy: KD	0.52	0.74	0.53	0.39	0.74	0.54
Hydropathy: GA	0.78	0.79	0.79	0.71	0.81	0.74
Statistical: TMHMM	**0.90**	**0.89**	**0.89**	**0.85**	**0.92**	**0.92**
Statistical: PHDhtm	0.83	0.84	0.82	0.78	0.73	0.85

The Matthew's correlation coefficients listed for the test sets are those of the average 200 best tables generated by the GA. Highest values are denoted in bold.

2.6 Testing the performance of the algorithm

The performance of the algorithm was measured using 3 different sets of 32 protein chains (8 membranous and 24 aqueous) that were not previously exposed to the training phase (Fig. 2, stage 12).

Each measurement was done in the following way: 200 best tables resulted from each training set out of the three, were selected. The tables which were different in their hydropathy as well as their window size and threshold values were exposed to the appropriate test set yielding a hydropahty analysis list which served for a creation of final result based on the most common helix positions.

The results of the optimized table were compared with those obtained by using the statistically based methods TMHMM and PHDhtm, and to the conventional hydrophobicity scales (KD and GES) (Table 2).

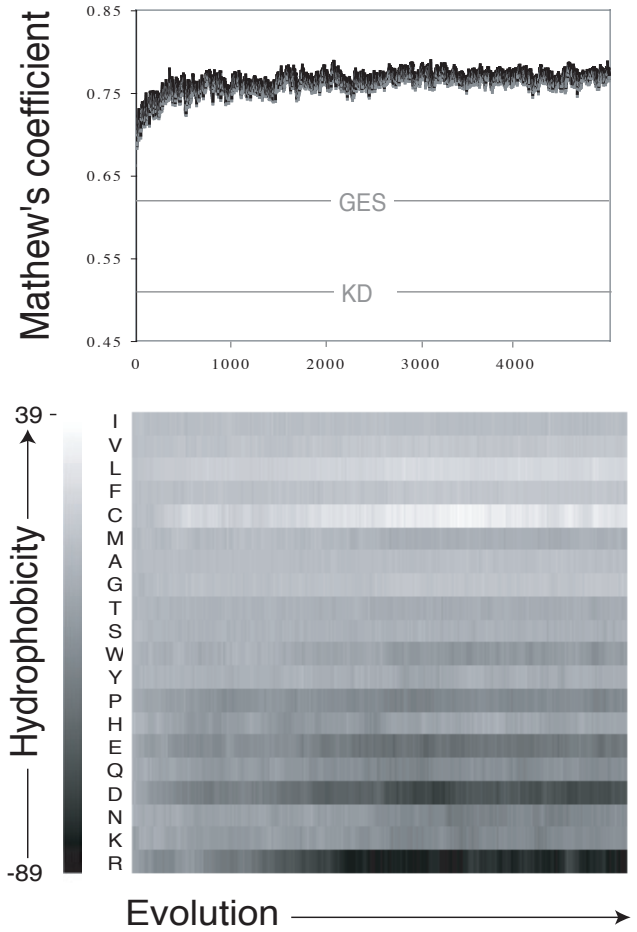

Fig. 4. Top panel: Predictive power of hydrophobicity tables generated by the GA from set 1 as a function of the evolution cycle during training. Bottom panel: Hydrophobicity table values during the course of the GA evolution whereby each column represents a single table and the color coding is according to the amino acid's hydrophobicity as noted in the color bar on the left (values are in kcal/mol). Initial values are on the left and final values are on the right.

After the evolution finished, the top 200 hydrophobicity tables were selected and used to obtain an averaged hydrophobicity profile of a particular protein. Explicitly, hydropathy analysis was used employing each of the 200 best tables. The results of these analyses were compared whereby a transmembrane segment was classified as true only if ≥80% of the tables predicted its presence. The final prediction was then used to calculate the Matthew's correlation coefficient as described above. The average values were then used to represent predictive power of the GA hydrophobicity tables.

3 RESULTS

Figure 4 (top panel) depicts the improvement in the predictive power of the hydropathy analysis on training set 1 as a function of the GA cycle number during the training phase. This improvement was found to be similar in the other training sets as well. Two different optimizations were performed, starting from the mating combinations of two different scales: GES and KD. These two matings exhibited similar behavior, manifested by a steep ascend of the predictive power in the first 1000 cycles, and by a steady state with a high predictive value.

Table 3. Amino acid hydrophobicity values (in kcal/mol) for the different scales used in the study

	KD	GES	GA Set 1	Set 2	Set 3
I	4.5	3.1	2.2	9.9	9.9
V	4.2	2.6	8.7	13.4	8.7
L	3.8	2.8	18.2	14.5	17.3
F	2.8	3.7	9.1	4.6	1.9
C	2.5	2.0	31.3	3.8	−3.8
M	1.9	3.4	−5.3	7.0	4.6
A	1.8	1.6	1.8	4.3	7.2
G	−0.4	1.0	7.7	0.1	−1.3
T	−0.7	1.2	−8.8	−11.4	−2.5
S	−0.8	0.6	−7.1	−0.3	−3.5
W	−0.9	1.2	−19.8	−4.2	3.6
Y	−1.3	−0.7	−9.4	−3.4	−18.4
P	−1.6	−0.2	−28.6	−35.3	−47.0
H	−3.2	−3.0	−12.2	−19.2	−23.0
E	−3.5	−8.2	−38.8	−49.4	−35.2
Q	−3.5	−4.1	−28.4	−29.6	−30.2
D	−3.5	−9.2	−66.8	−40.7	−55.1
N	−3.5	−4.8	−30.5	−52.5	−29.1
K	−3.9	−8.8	−25.0	−44.8	−41.2
R	−4.5	−12.3	−83.4	−64.5	−74.4

The final Matthew's correlation coefficients obtained from the training and the test set by using GES, KD, PHDhtm and TMHMM, in comparison with the GA average method, are given in Table 2. For example, the Matthew's correlation coefficient for the GES scale is 0.71 (test set 2), in contrast to the best table produced by the GA exhibiting a Matthew's correlation coefficient of 0.81. In addition, the predictive power of hydropathy analysis employing the optimized table on set 2 is found to be better than PHDhtm (Rost *et al.*, 1996) which manages to reach a Matthew's correlation coefficient of 0.73 only. The best table produced by the GA used an average window size of 20, and an average cutoff of 30 kcal/mol.

Figure 4 (bottom panel) depicts the evolution of the hydrophobicity values on training set 1 during the GA optimization. The optimized values are given numerically in Table 3, in comparison with those of other hydrophobicity scales.

In the first stages of the evolution process (∼1000–1500 generations), the hydrophobicity values undergo minor changes. Then, values of certain amino acids (Arg, Trp, Asp, Pro) undergo a significant change, after which the system enters a steady state in which the hydrophobicity values remain relatively stable.

Table 3 exhibits three groups of amino acids. (1) Amino acids which maintain the same hydrophobic or hydrophilic orientation in all hydrophobicity scales, where the GA-optimized table merely presents an amplification of the original values. More specifically, apolar amino acids (Ile, Val, Leu and Phe) become more hydrophobic, whereas polar amino acids (Arg, Lys, Asp, Glu, Asn, Gln, His, Pro and Tyr) become more hydrophilic. (2) Amino acids for which the hydrophobicity tables do not dictate a clear hydrophobic nature (Gly, Thr, Ser and Trp), and the GA-optimized table classifies them in the following manner: The hydroxylic amino acid Ser is mildly hydrophilic and the aromatic amino acid Trp is mildly hydrophobic (set 3). (3) Two amino acids, Cys and Met which

completely change their hydrophobic nature by the GA scale. In other words, in the GA-optimized table, Cys and Met are a hydrophilic amino acids (sets 3 and 1), and their value is negative, while in the other two tables it is hydrophobic and receives a positive value.

4 DISCUSSION

Our study presents an evolutionary algorithm concept developed in order to optimize currently available hydrophobicity tables and to enable better prediction of transmembrane segments in proteins.

4.1 The contribution of individual amino acids to the prediction of transmembrane helices

Analysis of the values that individual amino acids attain after optimization is revealing, in terms of the contribution of each amino acid towards the identification of transmembrane helices. The most obvious trend is that during the optimization process both the hydrophobicity and hydrophilicity values of the amino acids increase (Fig. 4). The effect is most pronounced in the basic amino acids Arg and Lys, consistent with the fact that their pKa values are farthest from neutrality and hence their de-ionization least probable.

4.2 Prediction accuracy

Hydropathy analysis was one of the first methods used to predict the presence of transmembrane helices. However, it has been virtually overridden by statistically based methods owing to their higher predictive value (Kall and Sonnhammer, 2002). Indeed, in the current study, hydropathy analyses using the two conventional hydrophobicity tables yielded Matthew's correlation coefficients which were much lower than prediction using TMHMM (0.39–0.74 versus 0.92, Table 2). Optimization of the hydropathy table using a GA, for 1000 iterations, shows an improvement in the prediction rate of the final optimized table over the conventional hydropathy indices. Given more evolution time, expressed by further iterations of the algorithm, the optimization ended with a final Matthew's correlation coefficient range from 0.71 (set 1) to 0.81 (set 2). Furthermore, optimization was performed using three different final test sets, so as not to over-optimize the resulting table to any specific sequence.

In the analysis of new genomes, there is reason to believe that hydropathy analysis based on the GA-optimized table may achieve predictions that do not change appreciably in varying genomes. This may contrast with pure statistically based algorithm that used a specified basis set to hone their accuracy. The GA we developed implements a heuristic approach to transmembrane helices prediction, that is based both on a set of known membrane proteins and on pure physicochemical principles that lie within the sequence level. The method did not use the positive inside rule or auxiliary genome information; yet, it achieved a high prediction value on the test set.

Previous studies have shown that nucleotide bias of genomes has a strong influence on the occurrence of hydrophobic amino acids found within transmembrane helices (Stevens and Arkin, 2000a). Therefore, this bias should be taken into account in methods that are developed to predict such regions. We believe that the method presented in this study and the optimized tables it produced, could be the base for further extensions, such as specific genome optimization of the hydrophobicity table, usage of the positive inside rule and propensity of amino acids to reside within transmembrane regions. In conclusion, we have shown that using a GA as a method for hydrophobicity tables optimization, is efficient in transmembrane segment prediction, and could be improved in order to increase both sensitivity and specificity of transmembrane detection methods.

ACKNOWLEDGEMENT

This work was supported in part by a grant from the Israel Science Foundation (784/01) to I.T.A.

REFERENCES

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–12.

Claros,M.G. and von Heijne,G. (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.*, **10**, 685–686.

Engelman,D.M. *et al.* (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.

Holland,J.H. (1975) *Adaptation in Natural and Artificial systems*. University of Michigan Press, Ann arbor, MI.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Holm,L. and Sander,C. (1997) Dali /FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Jones,D.T. *et al.* (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kall,L. and Sonnhammer,E.L. (2002) Reliability of transmembrane predictions in whole-genome data. *FEBS Lett.*, **532**, 415–418.

Koza,J.R. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* 1st, The MIT Press, Cambridge.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Rost,B. *et al.* (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.

Sonnhammer,E.L. *et al.* (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.

Stevens,T.J. and Arkin,I.T. (2000a) The effect of nucleotide bias upon the composition and prediction of transmembrane helices. *Protein Sci.*, **9**, 505–511.

Stevens,T.J. and Arkin,I.T. (2000b) Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins*, **39**, 417–420.

Tomita,M. and Marchesi,V.T. (1975) Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proc. Natl Acad. Sci. USA*, **72**, 2964–2968.

Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

von Heijne,G. (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, **341**, 456–458.

Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaen and eukaryotic organism. *Protein Sci.*, **7**, 1029–1038.

White,S.H. *et al.* (2001) How membranes shape protein structure. *J. Biol. Chem.*, **276**, 32395–32398.

Wimley,W.C. *et al.* (1996) Solvation energies of amino acid side chains and backbone in a family of host–guest pentapeptides. *Biochemistry*, **35**, 5109–5124.