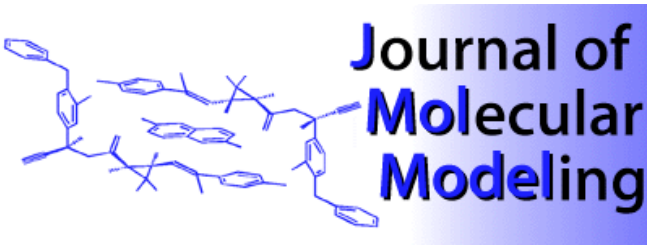
J. Mol. Model. **1999**, *5*, 317 – 324



FULL PAPER

Refined Genetic Algorithm Simulations to Model Proteins

Rainer König and Thomas Dandekar

EMBL, Postfach 102209, D-69012 Heidelberg, Germany. Tel: +49-(0)6221-387-372; Fax: +49-(0)6221-387-517;
E-mail: dandekar@embl-heidelberg.de

Received: 27 August 1999/ Accepted: 22 November 1999/ Published: 10 December 1999

Abstract The advent of completely sequenced genomes is leading to an unprecedented growth of sequence information while adequate structure information is often lacking. Genetic algorithm simulations have been refined and applied as a helpful tool for this question. Modified strategies are tested first on simple lattice protein models. This includes consideration of entropy (protein adjacent water shell) and improved search strategies (pioneer search +14%, systematic recombination +50% in search efficiency). Next, extension to grid free simulations of proteins in full main chain representation is examined. Our protein main chain simulations are further refined by independent criteria such as fitness per residue to judge predicted structures obtained at the end of a simulation. Protein families and protein interactions predicted from the complete *H. pylori* genomic sequence demonstrate how the full main chain simulations are then applied to model new protein sequences and protein families apparent from genome analysis.

Keywords Genomics, Entropy, Main chain, Structure criteria, *Helicobacter pylori*

Introduction

Genome and sequencing projects are yielding an exponentially growing amount of protein sequence data [1,2]. To understand their structure and function, protein modelling efforts have become an important tool [3]. Better protein structure and function prediction techniques are a focus of current research (reviewed in [4-6]).

For this challenge we apply and refine protein models using genetic algorithm simulations. We show recent improvements considering folding forces and search strategy in simplified models. Incorporation into full main chain models is possible. After further development these will allow a more detailed description of protein-protein and protein-solvent interactions. In fact, many protein interactions can be supported and new ones become apparent from genome comparisons of gene order, as we illustrate here for *Helicobacter pylori*.

However, for the high demand in three dimensional models of proteins from genome projects more immediate strategies are also desirable. For this, we improve grid free main chain genetic algorithm folding simulations by independent criteria to judge structure fitness at the end of the simulation and apply the full main chain protein simulations to genomics. A genetic algorithm model of the extracellular globular domain of an extended protein family revealed from genome analysis of *H. pylori* is shown as an example.

Part of this work was presented at the 13. Molecular Modeling Workshop, 25–26 May 1999, Darmstadt, Germany

Materials and methods

HP model

To investigate refined search strategies, genetic algorithm simulations were conducted in the context of the simplified HP model [7]. The HP model derives its name from the two types of residues which are only considered: polar/hydrophilic (P) and non-polar/hydrophobic (H). The protein folding space is simplified to a self avoiding random walk on a 2D or 3D square lattice. We use a simple energy scoring function [8] of minus one for any two hydrophobic residues directly contacting each other on the lattice (no diagonal contacts were counted).

Monte Carlo simulations

Monte Carlo simulations on entropy of the solvent were also tested in the context of the HP model. To study entropy effects, lattice spaces adjacent to the protein chain were modelled to be filled by solvent, called (small) water ensembles in the following. Long range interactions in the rest of the solvent (e.g. long range order brought about by solvating ions or exposed polar groups) were not considered in this simple model. Small water ensembles with two different properties surrounded the model protein: ordered and less ordered. The ordered water ensembles are adjacent to hydrophobic residues or the solvent (i.e. other water ensembles). The less ordered (with high entropy) exist if no hydrophobic residue is adjacent to them. The number of unordered water ensembles, N_i, was counted. It has been shown experimentally that hydrophobic molecules reduce the entropy of surrounding (aqueous) solvent [9]. The solvent entropy difference, S, between one protein chain conformation N_1 and a tested next one N_2 during the simulation was set to be proportional to the difference in the number of unordered adjacent lattice spaces counted. With the order parameter f (to be optimised) the probability for the new configuration to be chosen gets: $p \sim e^{(TS/T)} = f^{(N1-N2)}$. This implementation regards the entropy of the solvent according to Boltzmann statistics. The simple energy function from above is derived again (minus one for any hydrophobic contact), if the sum ("hydrogen bonds" in our model) of the connections of (water ensembles - water ensembles) and (water ensembles - hydrophilic residues) is counted and compared between two conformations.

The order parameter f allowed testing of different entropy weights during simulations, either alone or multiplying the f^{N1-N2} term with $e^{(-\Delta E/T)}$ yielding $p \sim e^{-F/T} = e^{(\Delta E - TS)/T}$ to consider also the energy difference ΔE between two chain conformations. F denotes the Helmholtz free energy (corresponding also to the Gibbs free energy if changes of volume and pressure are neglected). The new chain conformation was accepted if a random number between zero and 1/constant was smaller than $e^{-F/T}$.

The free energy function above was used with an artificial temperature T which was slowly decreased during the simulation, analogue to the simulated annealing method [10].

This model allowed examining of both entropy (ΔS, difference of ordered small water ensembles) and energy effects (ΔE; simplified in the context of the model; only hydrophobic interactions are considered or, alternatively, the effect of "hydrogen bonds").

Main chain protein folding simulations

These are grid free simulations. The protein main chain (N,C_α,C' and O) is modelled using internal coordinates and a set of seven standard conformations to assign ϕ and ψ values to the backbone [4]. The conformations of all residues along the amino acid sequence were successively collected together and decoded from a long bit-string (a "chromosome"). Starting from a population of random bit-strings, the quality of each encoded structure was judged by a fitness function composed of rewards and punishments. Five structure parameters, suitably weighted (further details in [11,12]) were calculated and summed up to judge structural fitness:

1) the total scatter of all (n) residue C_α-atoms (res), each (i) with (j) coordinates (x and y and z) around the common centre of mass (C_m), is considered (it is summed over all distances; the individual distance to the centre of mass for each residue, its radius of gyration, is calculated as the square root of the x, y and z component distance square):

$$\sum_{i=1}^{n} \sqrt{\sum_{j=x,y,z} \left(res_{ij} - C_{m_{ij}} \right)^2}$$

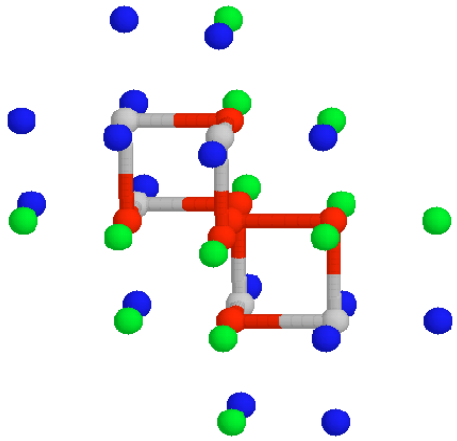


Figure 1 ***Peptide structure and solvent model on a three dimensional lattice.*** *Hydrophobic residues are shown in red, hydrophilic in grey. Adjacent solvent water molecules are represented as blue („unordered", directly adjacent to a hydrophilic residue of the protein chain) and green balls (more ordered, adjacent to a hydrophobic residue of the protein chain)*

2) distribution of hydrophobic residues only (M,I,L,V,Y,C,F and W; hydrophobicity scale according to [13]) around the centre of mass (same centre as in 1));

3) main chain van-der-Waals atomic overlaps;

4) conformational states which agree with the secondary structure (either known or predicted) for a given subsequence and

5) the selection for the formation of hydrogen bonds in ß-strands and ß-sheets and the formation of reverse turns in ß-hairpins.

Information from secondary structure prediction on helices and strands was utilised with residues found in such configurations kept fixed in appropriate standard conformations [14] during the simulation. All other residues were not fixed in the simulations, the genetic algorithm operated freely on all other residues.

Genetic algorithm simulation conditions

High quality bit-strings (after a random start) were selected preferentially as parents and mutated (one bit per string per generation) and recombined through cross-over (probability of recombination is 0.2 per bit string per generation and occurs at exactly one equivalent site chosen at random on each of the parental chromosome pairs) to yield the next parental generation of folds. A positive constant keeps the population of prediction trials richer since low fitness individuals may also survive. Simulations were run over many generations to allow convergence (the product of population and generation equalled at least 4 x 10^5, corresponding to a processing time for main chain simulation runs of 20 minutes on a VAX 7620 for a 46-residue protein). The best fold comparing the fittest individuals from ten runs with different random starts yielded the prediction in the full main chain simulations, at least 100 simulations were considered in the HP model simulations.

Test set

The following test set of small proteins with known three dimensional structure was used in the main chain folding simulations:

1IFM, 1PNH, 1PPT, 2OVO, 1MLI, 1EGL, 1HMD, 1GPT, 1EPR, 1DFN, 2CCY, 1CRO, 1CRN, 1TCG, 2CRD, 2BUS,

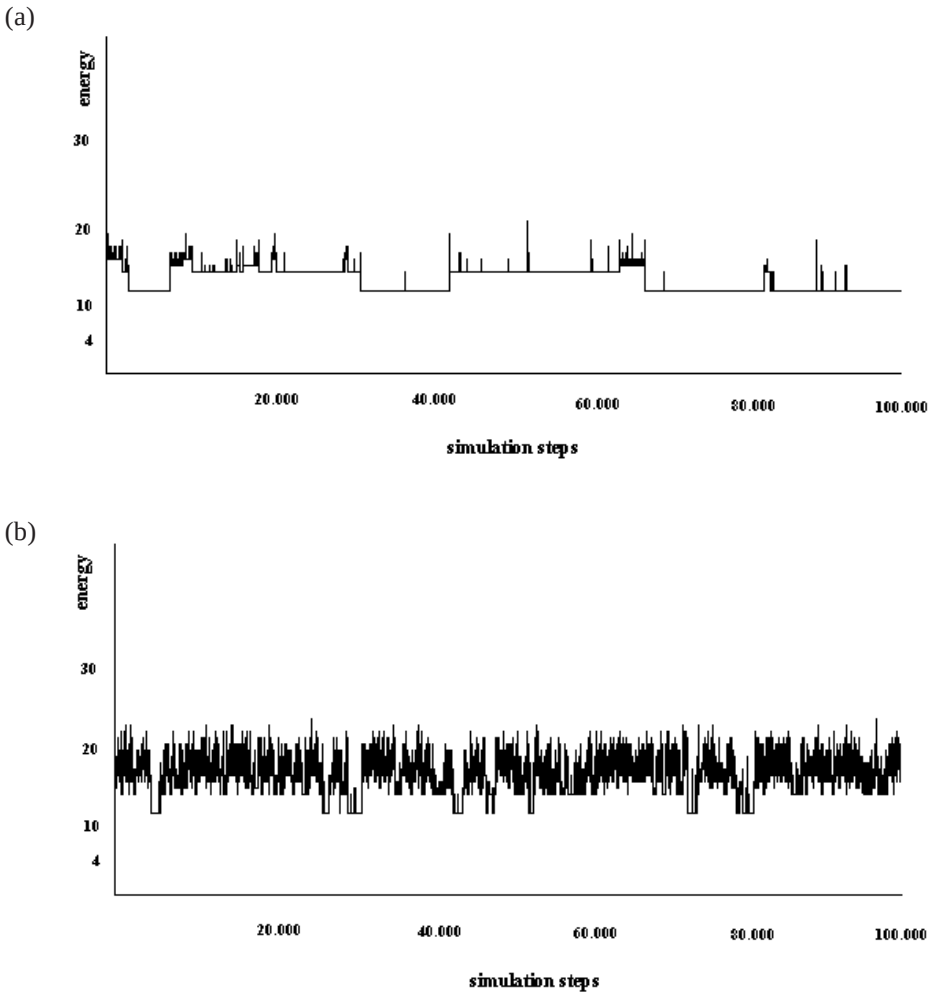


Figure 2 ***Entropy simulation.*** *Lattice simulations of a small protein chain (12 residues) are compared with a) a strong weight (order parameter f = 0.1) and b) a weak weight (order parameter f = 0.3) on entropy of the solvent. x-axis: Simulation step coordinate, each time 100000 steps were effected. y-axis: Energy values (number of hydrophobic contacts; calculated as described in Materials and Methods)*

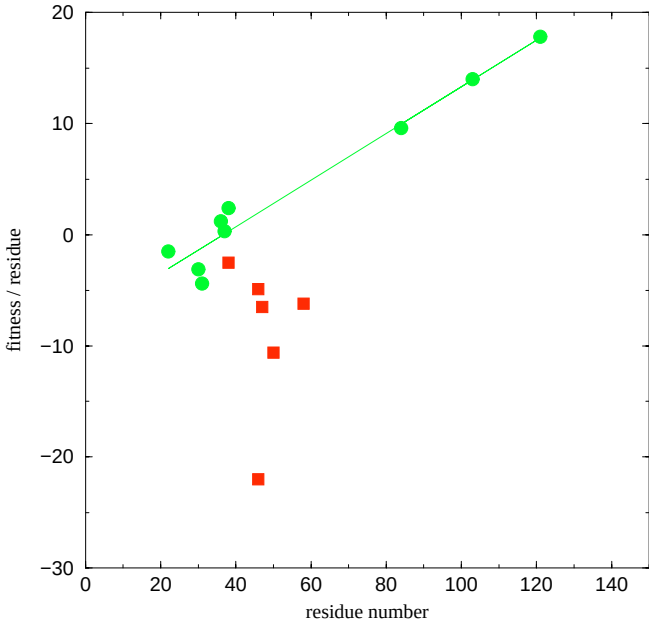


Figure 3 ***Fitness per residue in genetic algorithm protein folding simulations.*** *Simulations with good topology prediction and maximum RMSD of 5-6 Å are shown as filled green circles, the green line shows their regression line; failed simulations, violating one or both criteria, are shown as filled red squares. x-axis: protein size according to the number of residues. y-axis: normalized fitness per residue (fitness criteria as in Materials and methods)*

7PTI, 1BBI, 256B, 1ATX, 2GB1, all with different topology and known three dimensional structure.

Results and discussion

Refinement of protein models by representation of forces such as entropy

A more detailed modelling of folding forces such as entropic forces is desirable for protein models and to describe protein-protein interactions. This concerns for instance antigen-antibody contacts. Many more protein-protein complexes can be predicted from genome analysis by identification of conserved gene pairs (see application to genomics).

Simplified models we investigate use the HP model and a Monte Carlo algorithm to compare the results of different strengths of solvent entropic forces on a protein fold. The HP model folds a protein chain on a lattice, considering only two types of residues, hydrophobic (H) and hydrophilic (P) [7].

Figure 1 shows the native fold of a small model peptide studied. Red balls and sticks represent hydrophobic residues, grey ball-and-stick representation shows hydrophilic residues. The solvent water molecules are represented in small ensembles and shown as blue and green balls. Green balls represent more ordered water ensembles, which exist if hydrophobic residues are adjacent, whereas the blue balls represent less ordered water ensembles, not adjacent to any hydrophobic residue. The native fold of this small model structure is known.

Two model simulations (Figure 2) demonstrate different stability shown by the energy values obtained during successive simulation steps. They model two different parameter settings of the entropy. In Figure 2a the effect of strong entropic forces is modelled (order parameter f set to 0.1, see Materials and methods). The simulation (which in theese trials continues even if the global minimum has been found) stays most of the time in the global minimum, however, sometimes significant deviations and unfolding of the native structure are observed.

In contrast, setting the entropic forces low in the simulations (Figure 2b; order parameter f = 0.3), the global minimum is reached more slowly in the particular simulation

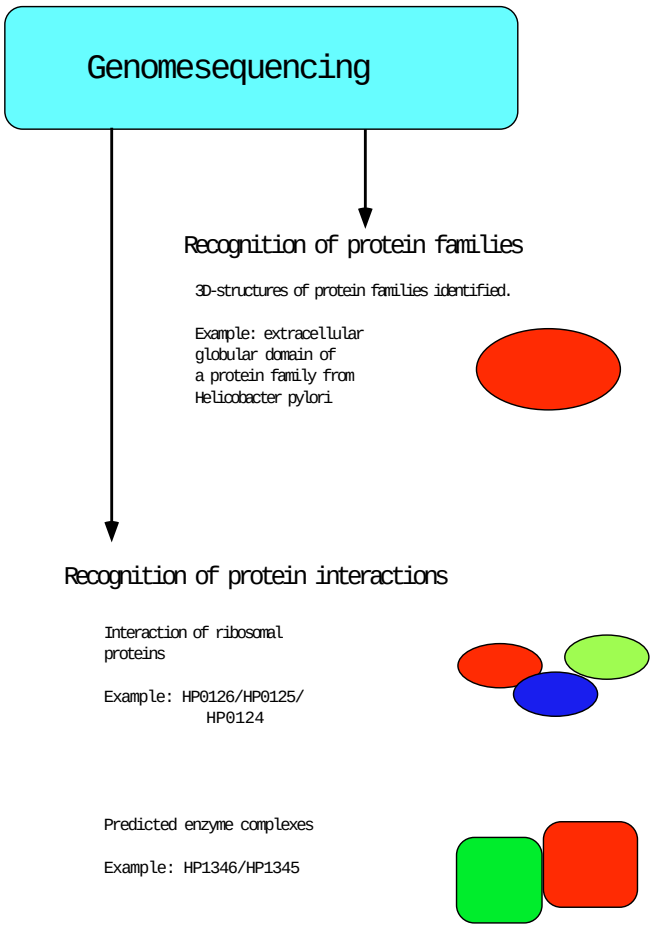


Figure 4 ***Genome analysis.*** *Some applications of our modelling efforts in genome analysis are sketched. This includes description of cytoplasmic protein domains in new recognized members of protein families (top) as well as models for the domains of interacting proteins (bottom)*

Scheme 1 *Pioneer search strategy*

Generation	**population** *(standard genetic algorithm)*	**population** *(genetic algorithm with pioneer search)*
i	B,B,B,B,A,B,B,C	B,B,B,B,A,B,B,C
i+1	B,B,D,B,B,B,B,B	K,L,M,H,I,J,N,O
i+2	B,D,E,B,B,B,B,B	R,Q,Z,Y,W,S,T,U

Each letter represents the exact bit string of one individual solution trial. The population with pioneer search creates every ten generations diverse new individuals whereas without it there is a tendency to stay around a (local) solution represented by individual B. Every ten generations, further new search space is explored („pioneered")

shown. However, the protein structure is more flexible and though it returns to the global minimum, on average the protein is often in a more unfolded state than in the simulations with strong entropy.

Even though simplistic, these and similar simulations are helpful to study the effect of different forces for protein folding such as the effect of solvent entropy, details of the hydrophobic packing, identification of the native state and protein stability. Furthermore, the strength of both entropic and hydrophobic forces are easily quantified including time spent in different folded and unfolded states for a specific chain type. For full main chain simulations using the genetic algorithm (see Materials and Methods) the neighbouring water molecules will next be approximated by a water shell around the protein. Different weights of the entropic forces as described and tested above can directly be implemented as an additive term to the fitness function and next be optimised as an additional selection criterion. The incorporation of entropic forces to model protein interactions should allow a more detailed examination and modelling of protein interactions such as revealed by genome analysis of *H. pylori* (see application to genomics).

Refinement of the genetic algorithm search strategy by pioneer search and systematic recombination

Apart from a more detailed representation of folding forces, improved simulation results may be expected from a refined search strategy. This is again conveniently tested in the HP model as many simulations can be run quickly and compared. One challenge for genetic algorithm simulations is to keep the population of solution trials sufficiently rich and diverse [15,16]. Motivated by concepts such as Taboo search [17], we examined different modifications of the genetic algorithm strategy. After investigating a number of alternative strategies, two search strategies yielded an improvement in search performance (see schemes 1 and 2).

One strategy, "pioneer search" [8] explores new regions of the search space forming "pioneering" individuals if the population is not rich enough (Scheme 1). Every ten generations the newly created individuals are tested to see if they differ from *every* individual of the parent population. Individuals are discarded if they do not differ. In contrast to taboo search [17], this strategy seems only to have a short term memory as only the one generation back is considered to form new individuals to explore new search space. However, these parent individuals represent by their survival a long term memory from the past. The condition to be at least in one detail different from the whole previous population pushes forward into new search space while keeping longer memory in those individuals which are only minimally different from the previous population. In our simulations this was sufficient to prevent oscillations between reoccurring populations. The strategy is now examined further. It yields a gain in searching for different HP-chains, e.g. 14% less evaluations to find the global minimum for a 20 residue chain on a lattice.

An additional improvement of the search performance was achieved by systematic recombination. As the recombination event itself is random in our standard genetic algorithm procedures, systematic recombination of individuals may potentially improve the evolution. The effect of systematic recombination was also examined in simplified protein folding trials (Scheme 2). In an extensive comparison for a number of different protein chains in the context of the simplified HP model, this strategy gave a speed up in search speed (a factor 3/2 for a 20 residue chain) and identified the global mini-

Scheme 2 *Systematic recombination strategy*

1. Identify the best individual from the parent generation
2. Choose another individual (with linear probability to be picked according to fitness; the individual has to differ from the best individual)
3. Do systematic cross-over at all possible cross-over points between the two individuals
4. Pick the best recombination result (in terms of fitness)
5. Repeat several times and add these individuals to the population

Table 1 *Conserved gene pairs (examples) found in genome comparisons [a]*

Triple genome comparisons		
Methanobacterium Thermoautotrophicum (MT)	*Pyrococcus horikoschii (PH)*	*Bacillus subtilis(BS)*
MT0007/0008	PH1772/1773	BS0121/0122 [b]
MT0013/0014	PH1767/1768	BS0126/0127
MT0018/0019	PH1764/1763	BS0130/0131
MT1055/1056	PH1541/1542	BS0110/0111
Helicobacter Pylori (HP)	*Mycoplasma Pneumoniae (MP)*	*Haemophilus Influenzae (HI)*
HP0126/0125/0124	MP037/038/039	HI1320/1319/1318 [c]
HP0402/0403	MP048/MP049	HI1312/1311
HP1198	MP326/MP327	HI0515/0514
Two genome comparison		
Helicobacter Pylori (HP)	*Mycoplasma Pneumoniae (MP)*	
HP1346/1345	MP411/MP412 [d]	

[a] The genome number identifiers indicate the position of each gene.
[b] All these pairs are conserved ribosomal protein genes.
[c] The triple is found for ribosomal proteins. The two subunits for phenyl tRNA synthetase are the next example, DNA-directed RNA polymerase beta and beta' subunit is the last. The RNA polymerase is fused to one protein in HP.
[d] The example (glyceraldehyde 3-phosphate dehydrogenase; phosphoglycerate kinase) is not conserved in HI; however, with lower confidence also this conserved pair suggests a direct interaction (e.g. multienzyme complex) between these two consecutive enzymes of glycolysis. This would be biological meaningful (an adaptational advantage) and can now be tested experimentally

mum more reliably than the simple genetic algorithm alone [8].

Current investigations in protein folding trials now with grid free representation of the protein main chain (see Materials and Methods) indicate that particular in situations where there are not many different solution trials in the population (because the population has evolved for many generations), and in a final overall competition of few best solutions the systematic recombination is advantageous.

Judging the quality of the models obtained

In contrast to the strategies in (i) and (ii) which may lead to a decisive improvement in protein modelling in future, the quality of the final best structure obtained by the genetic algorithm simulations is an independent and readily applicable method to test and refine protein predictions from the genetic algorithm. In test trials for this, our full main chain protein folding simulation were used from the start, applying our genetic algorithm fitness criteria and a test battery of several small proteins with experimentally resolved structure (see Materials and Methods).

The RMSD of the simulated structures to the observed structure (considering all main chain atoms) was compared with the fitness value per residue (Figure 3). This criterion may be a useful tool to differentiate between successful and less successful simulation runs in the context of our full main chain models. On the y-axis a normalised fitness value per residue is given, on the x-axis the number of residues in each protein structure is plotted. For simulations (filled green circles) which were found to have good topology predictions and RMSDs not more than 5-6 Å, the normalised fitness value per residue increased linearly with the size of the structure modelled in these simulations. For comparison, simulations with higher RMSDs and bad topology predictions are shown (filled red squares). These trials suggest that the fitness values per residue in the unsuccessful simulations is lower than that achieved in the successful trials. This will be studied further covering more trials and studying more protein folds. Similarly to several other alternative criteria in judging correctness of predicted structures, this quantitative measure will be quite helpful to judge structure quality in blind prediction trials.

Application to genomics

In several of the genomes we are investigating, species of specific protein families become apparent by full genomic

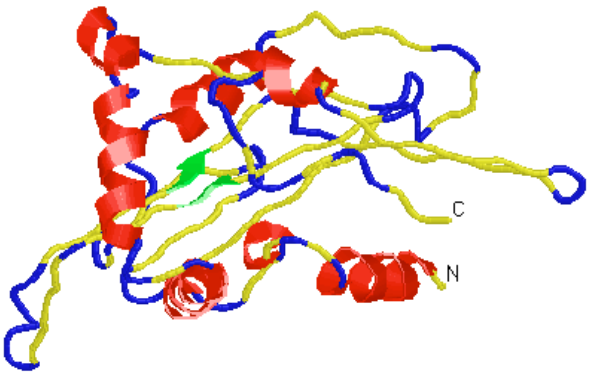


Figure 5 ***Model of extracellular globular domain from H.pylori protein HP1525.*** *The C_α-chain trace was predicted by the genetic algorithm and is shown in cartoon representation. Around a central sheet region (yellow; well packed central two strands marked in addition in green) the helices (red) are packed. Coil and turn regions are in blue. C- and N-terminus are indicated*

comparisons using sequence similarity (Figure 4). This concerns for instance a family of outer membrane proteins in *H. pylori* [18] including the proteins HP0079, HP1156, HP0472, HP0373 and HP1525 (HP genome identifiers are given). These proteins have no homologous sequence with known 3D structure, no characterised SMART or PFAM domains and no transmembrane domains. However, for the non-transmembrane, extracellular globular domains of these proteins, we can apply a full main chain genetic algorithm simulation to get a first picture of this not yet well characterised protein family. The genetic algorithm based model for the globular, extracellular domain of the *H. pylori* protein HP1525 is shown in Figure 5. Around a central sheet region the helices are packed, C- and N-terminus are indicated. This *ab initio* model, based on sequence and secondary structure prediction, may still have a high error in some regions (higher than 5-6 Å RMSD). However, the complete main chain topology prediction provides a starting point for experiments such as probing and comparing predicted surface epitopes with available antibodies. In particular, the relationship to *H. pylori* iron(III) dicitrate transports (HP0807, HP0686, HP0608) suggested previously [18] can now be examined by suitable experiments. Residues found in close contact to each other as well as additional experimental data can be applied directly as a selection criterion for a refined model. Additional data from such experiments can easily be incorporated into the simulation with appropriate selection weights (details on weights in [19]).

Besides new protein families, new protein interactions are revealed by our comparative genome analysis. Conserved consecutive gene order as judged by the homologous proteins encoded in the same consecutive order found in several different prokaryotic genomes is indicative of a direct physical interaction of these encoded proteins [20]. Inclusion of entropic forces (see above) with appropriate weights will help to refine our genetic algorithm models regarding such protein interactions. Several conserved gene pairs identified in two new triple genome comparisons involving arachaea (top) and *H. pylori* (middle) are shown in Table 1. The table shows that conserved gene pairs are still present even if one species included in the triple comparison is only very distantly related to the other two, demanding a high pressure for evolutionary conservation of gene order of this particular genes.

Far more gene pairs can be found to be conserved in genome order if only two species are considered. An interesting example for *H. pylori* is given in the bottom of Table 1. As in such a case the criteria for evolutionary conservation are less strict the prediction for a physical interaction of the encoded proteins becomes less certain. The biological context of the example shown is another independent indication, two consecutive enzymes of a metabolic pathway would advantageously interact. This will be tested further including protein models for the two *H. pylori* enzymes.

Conclusion

The current era of genome sequencing and genomics creates an increasing demand for three dimensional predictions from sequence [3-6]. Complete genome sequences for important pathogens such as *H. pylori* are now available. For these challenges we refine genetic algorithm modelling search strategies (systematic recombination, pioneer search) and investigate the inclusion of entropic forces to better model protein interactions. Our refinements are successful in the context of the simplified HP model and are currently transferred to full main chain models. The full main chain protein models can be improved further by independent structure judgement criteria such as the fitness per residue achieved at the end of the simulation. The main chain topology models are applied to study new protein interactions and protein families apparent from genome analysis such as *H. pylori* but also other currently sequenced genomes.

Acknowledgements We thank Sean O'Donoghue and Louise Kelly for discussion and comments. This research was generously supported by DFG and BMBF.

References

1. Marshall, E. *Science* **1999**, *284*, 1906-1909.
2. Rousley, S.; Lin, X.; Kechum, K.A. *Curr. Opin. Plant Biol.* **1998**, *1*, 136-141.
3. Teichmann, S.A.; Chothia, C.; Gerstein, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 390-399.
4. Fischer, D.; Eisenberg, D. *Curr. Op. Struc. Biol.* **1999**, *9*, 208-211.
5. Shortle, D. *Curr. Biol.* **1999**, *9*, R205-209.
6. Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M.A.;Yuan, Y.P. *J.Mol.Biol.* **1998**, *283*, 707-725.

7. Lau, K.F.; Dill, K.A. *Macromolecules* **1989**, *22*, 3986-3997.
8. König, R.; Dandekar, T. *Biosystems* **1999**, *50*, 17-25.
9. Kauzmann, W. *Adv. Prot. Chem.* **1959**, *14*, 1-63.
10. Kirkpatrick, S.; Gelatt, C.D; Vecchi, M.P. *Science* **1983**, *220*, 671-680.
11. Dandekar, T.; Argos, P. *J. Mol. Biol.* **1994**, *236*, 844-861.
12. Dandekar, T.; Argos, P. *J. Mol. Biol.* **1996**, *256*, 645-660.
13. Manavalan, P.; Ponnuswamy, P.K. *Nature* **1978**, *275*, 673-674.
14. Rooman, M.J.; Kocher, J.P.; Wodak, S.J. *J. Mol. Biol.* **1991**, *221*, 961-979.
15. Burke, E.K.; Newall, J.P.; Weare, R.F. E*vol. Comput.* **1998**, *6*, 81-103.
16. Clark, D.E.; Westhead, D.R. *J. Comput. Aided Mol. Des.* **1996,** *10*, 337-358.
17. Cvijovic, D.; Klinowski, J. *Science* **1995**, *267*, 664-666.
18. Huynen, M.; Dandekar, T.; Bork, P. *FEBS Letters* **1998**, *426*, 1-5.
19. Dandekar, T.; Argos, P. *Protein Engineering* **1997**, *10*, 877-893.
20. Dandekar, T.; Snel, B.; Huynen, M.A.; Bork, P. *Trends Biochem. Sci.* **1998**, *23*, 324-328.