

# Consistent sets of secondary structures in proteins

Piotr Berman<sup>1</sup> and Jieun Jeong<sup>1</sup>

<sup>1</sup>The Pennsylvania State University {berman, jijeong}@cse.psu.edu

**Abstract.** Ab initio predictions of secondary structures in proteins have to combine local predictions, based on short fragments of the protein sequence, with consistency restrictions, as not all locally plausible predictions may be simultaneously true.

We use the fact that secondary structures are patterns of hydrogen bonds and that a single residue can participate in hydrogen bonds of at most one secondary structure. Consistency of fixed-sized pieces of secondary structures is the easiest to approximate and we formalize it as 1-2 matching problem. Consistency of entire secondary structures is a version of set packing. We also investigate how to form a simple problem if we add the requirement that the secondary structure and the loops that connect them fit together in a metric space.

Every problem that we investigated is MAX-SNP hard and it has a constant factor approximation. Computational experience suggests that in biological instances, we can find nearly optimal solutions using heuristics.

## 1 Introduction

One of the goals of bio-informatics is finding a way to predict the shape of a protein based on its sequence of residues (*i.e.*, the amino acids). The first stage of the shape prediction of a protein often consists of the identification of the set of its secondary structures (see the introduction of [6]).

The primary structure of a protein is its chain of *amino acids*, also called *residues*, that are connected by covalent bonds. There are 20 different possible residues (with rare exceptions) and thus a protein can be described as a sequence of letters from an alphabet of size 20. A protein chain created by the transcription process *folds* into its eventual shape. A large part of a typical protein consists of periodic *secondary structures*,  $\alpha$ -helices (see Fig. 1) and  $\beta$ -sheets of two kinds: parallel and anti-parallel (see Fig. 2).

There exist a number of methods (*e.g.*, Rost [17]) to provide *local predictions* of the form “residue  $i$  belongs to a  $\beta$ -sheet/ $\alpha$ -helix/none” with a certain degree of confidence. However, there is a difference between knowing which residues belong to what kind of structures, and identifying those structures. This problem does not arise for  $\alpha$ -helices because an  $\alpha$ -helix consists of a single fragment of a protein chain that is folded in a helical pattern. On the other hand, a  $\beta$ -sheet is formed from two relatively straight chain fragments that are connected with a pattern of hydrogen bonds. Thus there are two parts to predicting  $\beta$ -sheets: predict the chain fragments that they include (which can be called  $\beta$ -strands)

and the contacts between these strands. The problem discussed in this paper is predicting the set of  $\alpha$ -helices and  $\beta$ -sheets (both the strands and the contacts).

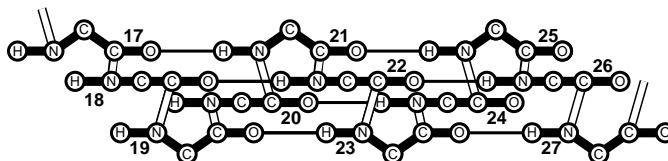
Several researchers proposed methods to predict sets of  $\alpha$ -helices and  $\beta$ -sheets (contacting pairs of  $\beta$ -strands) by forming a *consistent set* of individual predictions. Cheng and Baldi [6] used a neural network to provide confidence levels for small fragments of  $\beta$ -sheets, and they formulated a number of combinatorial consistency rules. In the first phase of their algorithm they select  $\alpha$ -helices and  $\beta$ -strands, and in the second they select the contacts between the strands in a greedy manner. The earlier work of Zhu and Braun [20] dealt with a similar problem, although their global optimization algorithm was rather implicit. Berger et al. (*e.g.*, [14]) considered a *recognition problem* of certain secondary structures in a similar framework — local propensity/likelihood measures and assembling a global solution.

The local measures can be obtained in a number of ways. Cheng and Baldi [6] used a neural network to provide confidence levels for their local predictions, Zhu and Braun [20], as well as Berger et al. [14], used simple formulae applied to the statistics of protein structures that are known experimentally [4, 5]. This approach was used much earlier by Hubbard [9, 10]. New studies on statistic properties of  $\beta$ -sheets can improve the accuracy of such predictions [7].

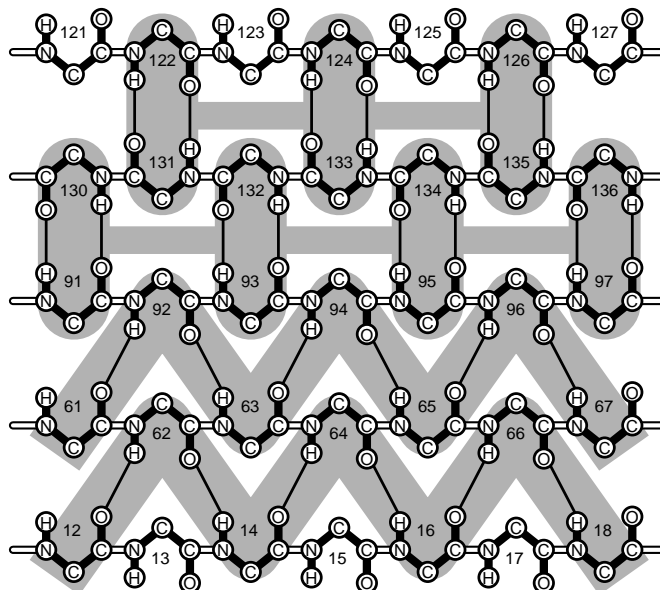
This suggests that one can consider the prediction of secondary structures as a global optimization problem: define a *domain* (local predictions), *feasible solutions* (sets that satisfy consistency rules) and an *objective function* (sum of confidence levels). It is far from clear which optimization problem is good in this context. On one hand, we want to exclude incorrect solutions; on the other hand, we want to have a relatively simple problem that can be solved in reasonable time. The most rigorous consistency test is that there exists a placement of the atoms in 3D space that satisfies a number of physical and chemical laws and which validates our local predictions. Meiler and Baker [13] use local predictions to extend the catalog of possibilities used by ROSETTA program [19]. This approach is often very effective, however, at the core, it is a simulated annealing method and it may be very important to have a reasonably good starting conjecture. Thus, we need simpler heuristics to provide such conjectures.

We investigated global optimization problems that can be used to find the most likely consistent set of “local conjectures”. By formulating the condition of consistency in several ways, we obtain three problems: maximum weight 1-2 matching, maximum weight packing of intervals split into two equal parts, and maximum packing of protein structures in which every two structures are metrically consistent.

In the maximum weight 1-2 matching our domain consists of very short fragments of  $\alpha$ -helices and  $\beta$ -sheets, and consistency condition is that the same residue cannot be in a  $\alpha$ -helix and in another secondary structure, and that it cannot be in more than two  $\beta$ -sheets. We show that this problem cannot be solved exactly in polynomial time, and that it can be approximated as well as the problem of maximum length TSP (a.k.a. taxi ripoff problem). In maximum weight packing the secondary structures are represented as sets of individual



**Fig. 1.** A schematic view of an  $\alpha$ -helix (backbone atoms only). Bonds inside amino acids are black, peptide bonds are white and hydrogen bonds are thin.



**Fig. 2.** Characteristic sets of  $\beta$ -sheets are indicated by gray outlines.

predictions, and our contribution is to structure these sets in such a way that the resulting packing problem is easy to approximate (unrestricted set packing problem is one of the hardest problems to approximate, [8]).

Lastly, we impose an additional rule on the solutions to the set packing problem, to reflect the fact that the secondary structures have to be placed in a metric space. We obtained a simple condition that would be as stringent as possible, and still leave a problem that is easy to approximate.

## 2 Abstract representation of secondary structures

A secondary structure has a specific shape that is enforced by its set of hydrogen bonds. Together with the bonds of the protein backbone (bonds inside amino acids and the peptide bonds that bind amino acids together), these bonds form periodic structures that are rigid in two dimensions ( $\beta$ -sheets) or in three dimensions ( $\alpha$ -helices). There exists a number of other *secondary structures*, but

from the point of view of the problems we formulate they can be treated like very short  $\alpha$ -helices.

When we model consistency of a set of structures with set packing we represent a structure as its *characteristic set* that consists of numbers of the residues that participate in hydrogen bonds of a respective structure.

The simplest structure to consider is an  $\alpha$ -helix which can be specified by two integers  $b, e$  and which has hydrogen bonds between pairs of the form  $(j, j + 4)$  for  $b \leq j, j + 4 \leq e$  (see Fig. 1 for an example).

Let  $par(i) \stackrel{\text{def}}{=} i \bmod 2$ . The second type of structure is an anti-parallel  $\beta$ -sheet which can be specified by 4 integers  $b_0, e_0, b_1, e_1$  where  $par(b_0) = par(e_0)$  and  $e_0 - b_0 = e_1 - b_1$  and which has hydrogen bonds between elements of pairs of the form  $(i, j)$  such that  $b_0 \leq i \leq e_0, b_1 \leq j \leq e_1, par(i) = par(e_0)$  and  $i + j = b_0 + e_1$ .

The third type of structure is a parallel  $\beta$ -sheet which can also be specified by 4 integers  $b_0, e_0, b_1, e_1$  where  $par(b_0) = par(e_0)$  and  $e_0 - b_0 = e_1 - b_1 + c$  where  $c \in \{0, 2\}$  and which has hydrogen bonds between elements of pairs of the form  $(i, j)$  such that  $b_0 \leq i \leq e_0, b_1 \leq j \leq e_1, par(i) = par(e_0)$  and  $i - j \in \{b_0 - b_1, b_0 - b_1 + 2\}$ .

We re-order the set of residue numbers of a protein, say  $1, 2, \dots, n$  to segregate the numbers according to their parity,  $(1, 3, \dots, n, 0, 2, \dots, n - 1)$  (we assume w.l.o.g. that  $n$  is odd). Each characteristic set of a structure that we have implicitly defined is a union of two contiguous intervals in that ordering.

For an  $\alpha$ -helix specified by  $b, e$  these are intervals  $\{2i : b \leq 2i \leq e\}$  and  $\{2i + 1 : b \leq 2i + 1 \leq e\}$ . For example, an  $\alpha$ -helix specified by 4, 20 has intervals  $\{4, 6, \dots, 20\}$  and  $\{5, 7, \dots, 19\}$ .

For a  $\beta$ -sheet specified by  $b_0, e_0, b_1, e_1$  these intervals are  $\{b_0 + 2i : b_0 \leq b_0 + 2i \leq e_0\}$  and  $\{b_1 + 2i : b_1 \leq b_1 + 2i \leq e_1\}$ . For example, a parallel  $\beta$ -sheet specified by 50, 60, 21, 29 will have intervals  $\{50, 52, \dots, 60\}$  and  $\{21, 23, \dots, 29\}$ .

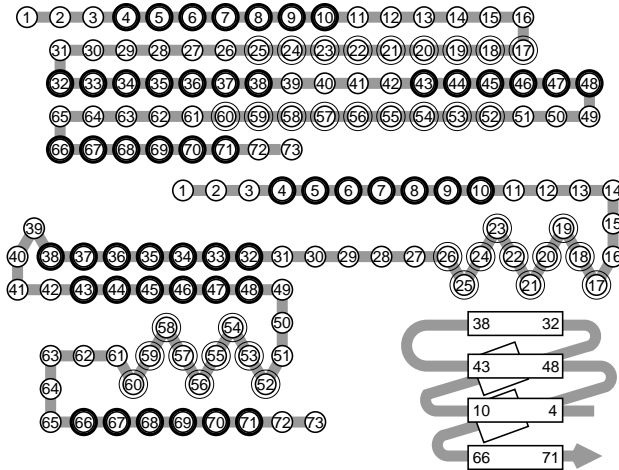
Note that  $\beta$ -sheets may form consistent (disjoint) sets in which amino acids participating in two  $\beta$ -sheets alternate, indeed, in the case of actual proteins, this happens very frequently.

Fig. 3 shows an example of the primary and secondary structures of a protein.

### 3 Maximum weight 1-2 matching

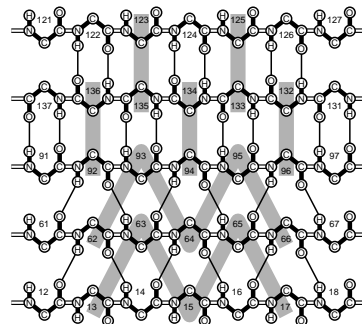
A possible method of predicting secondary structures would make predictions at the level of individual hydrogen bonds, or, to be more selective, at the level of groups of 3 to 4 adjacent hydrogen bonds.

Residues of such a group form two triples, Tr-pair for short, say  $(i - 1, i, i + 1)$  and  $(j - 1, j, j + 1)$ . In an  $\alpha$ -helix they could contain three hydrogen bonds,  $(i - 1, j - 1), (i, j)$  and  $(i + 1, j + 1)$ ; in a parallel  $\beta$ -sheet, they could contain three bonds  $(i - 1, j - 1), (j - 1, i + 1)$  and  $(i + 1, j + 1)$ , while in an anti-parallel  $\beta$ -sheet they could contain (double) bonds  $(i - 1, j - 1)$  and  $(i + 1, j + 1)$ . One can describe a structure as a set of its Tr-pairs. In this way a set of possible secondary structures defines a graph: consecutive triples contained in structures form nodes, and the Tr-pairs form edges.



**Fig. 3.** An example of secondary structures in a protein (Baker yeast ATX1 metal-lochaperone). Residues of  $\beta$ -strands are marked with thicker circles, and  $\alpha$ -helices are marked with double circles. First we have an unfolded chain with the structures being marked. Next we show how the  $\alpha$ -helices fold and one of the  $\beta$ -sheets and finally we show the experimentally confirmed arrangement of the secondary structures.

Note that while two different Tr-pairs from experimental structures of a protein may overlap, the overlaps of the centers are very restricted: a residue can either be the center of a triple that belongs to two Tr-pairs of an  $\alpha$ -helix, or of a triple that belongs to two Tr-pairs of a parallel  $\beta$ -sheet; however, a triple from an anti-parallel  $\beta$ -sheet can belong to one Tr-pair only. Thus we give edges *width* 1 and 2 and add a condition that the sum of width of selected edges adjacent to a single node is at most 2. We will call sets satisfying that condition *1-2 matchings*. For appropriately defined edge weights our task will be to find a maximum weight 1-2 matching.



**Fig. 4.** 1-2 matching corresponding to the example in Fig. 2

Notice that in solutions that correspond to actual protein structures we cannot have cycles. Therefore we redefine the maximum weight 1-2 matching problems to allow only the cycle-free solutions. If we have width 1 edges only, we would get a problem of finding a maximum weight cycle-free 2-matching, or, equivalently, a maximum weight Traveling Salesman Tour (we credit the steps that follow the edges with the weight of these edges, and steps that do not follow any edge we credit with 0). An efficient algorithm for a maximum weight TSP was provided by Serdyukov [18] and it has (nearly best) approximation ratio  $3/4$ .

The cycles of width 1 edges can be created if we have conflicting prediction of parallel  $\beta$ -sheets. Many proteins have very few parallel  $\beta$ -sheets, or none at all, while edges defined by  $\alpha$ -helices cannot form cycles (they form chains of the form  $(i, i + 4, i + 8, \dots)$ ). Thus one may ask if we can solve the maximum weight 1-2 matching problem in graphs in which width 1 edges form no cycles. However, even this restricted version of 1-2 matching is MAX-SNP hard (proof in the full version, see Appendix A).

Our approximation for 1-2 Matching is the following. First, we find an approximate maximum weight TSP tour that uses edges of width 1 using Serdyukov's algorithm (or a maximum weight 2-matching if we have no cycles of width 1 edges), this is a solution  $M_1$ . Second, we find a maximum weight matching that consists of edges of width 1 and 2, this is a solution  $M_2$ . We pick the better of these two solutions. The resulting approximation ratio is  $3/5$  in the general case and  $2/3$  if we have no width 1 cycles. (Proof in the full version).

A weakness of this method is that a matching does not have to correspond to an arrangement that is possible in three dimensions. This motivates formulation of other optimization problems which may be hard to solve exactly, but which have fewer solutions that are formally valid, but that do not correspond to possible arrangements of structures.

## 4 Set packing

As we have discussed, we can predict whole secondary structures using local methods. Rather than encoding these predictions as weights of corresponding edges, we can view them as jig-saw puzzle pieces and try to assemble them together. More precisely, we seek a consistent set of pieces with as large joint weight as possible.

The simplest kind of consistency of secondary structures that we consider is the disjointness of their characteristic sets (see Section 2). This is a set packing problem, and the quality of the available approximation algorithm is determined by the properties of the family of these sets.

Set packing is a special case of independent set problem: we form a graph in which nodes are sets from our family, and edges are overlaps, *i.e.*, pairs  $\{A, B\}$  such that  $A \cap B \neq \emptyset$ . While this problem is very hard to approximate in general, for graphs with special properties we have much better polynomial time approximation algorithms.

The first salient property is that each set is represented as a pair of intervals on the same line (we reorder the integers to have all even numbers first and all odd numbers later). This allows to use the *fractional local ratio* algorithm of Bar-Yehuda et al. [1] that works in polynomial time and guarantees approximation ratio 4 (*i.e.*, it guarantees to find a solution with weight that is at most 4 times smaller than the optimal one). However, this is a relatively slow algorithm: for every set that is being considered we may have to solve a linear programming problem.

We can establish another graph property that we will call conflict number. *Independence number*  $\iota(A)$  of a set of nodes is the maximum size of an indepen-

dent set in  $A$ ; *restricted neighborhood* of a node,  $N(u, A)$ , is the set of nodes of  $A$  that are connected to  $u$ , and  $u$  itself; *conflict number*  $C(G)$  of a graph is

$$\max_{A \subset V(G)} \min_{u \in A} \iota(N(u, A)).$$

One can use the algorithm of Berman and DasGupta [2] to assure approximation ratio equal to the conflict number. This algorithm runs as fast as a greedy algorithm.

The sets in our families are peculiar pairs of intervals: they differ in size by at most 1; we will call them *pairs of almost equal integer intervals*.

**Lemma 1.** *The conflict number of a family of pairs almost equal integer intervals is not larger than 4.*

**Proof.** In the full version (see Appendix C). □

**Lemma 2.** *Given a graph  $G$  with conflict number  $C$ , let  $W(G)$  be the maximum weight of an independent set. In polynomial time we can find an independent set with weight at least  $W(G)/C$ .*

**Proof.** We assume that  $C$  is a constant. In polynomial time we can construct an induced subgraph  $G'$  with possibly altered weights such that for some  $w$  (a)  $W(G') \geq W(G) - Cw$ , and (b) an independent set of  $F'$  with weight  $x$  can be modified into an independent set of for  $G$  with weight  $x + w$ .

First we find a node  $u$  with conflict number  $C$  and we record  $w = w(u)$ . Next, for every  $v \in N(u, V)$  we subtract  $w$  from  $w(v)$ . Finally, we remove nodes with non-positive weights from  $V$ .

Once we have a solution for this new instance, we check if it contains any node from  $N(u, V)$ . If yes, we add  $w$  to the weight of each such node. If not, we restore the weight of  $u$  and insert  $u$  to the solution. □

the algorithm from Lemma 2 yields the following

**Theorem 1.** *Given a family of  $n$  pairs of almost equal integer intervals with weights, such that the maximum weight set packing has weight  $W$ , in time  $O(n^2)$  we can find a set packing with weight at least  $W/4$ .*

## 5 Metric consistency

Secondary structures have a certain degree of rigidity, and this allows us to predict that certain sets of secondary structures cannot co-exist even though their characteristic sets are disjoint. We will use an abstract definition of rigidity.

We say that subgraph  $H$  of  $G$  is *rigid* if distances (shortest path lengths) computed inside  $H$  are equal to distances between nodes of  $H$  that are computed in  $G$ . For example, suppose that in a graph  $G$  we have a path  $(u_0, \dots, u_k)$  of edges of length 1; we say that this path is rigid in  $G$  if for  $0 \leq i < j \leq k$  the shortest path from  $u_i$  to  $u_j$  has length  $j - i$ . Testing the rigidity of a subgraph is very simple, so we have a simple test of consistency of structures that is significantly more realistic than simple disjointness of characteristic sets.

	$\alpha$	$\beta$	$\alpha/\beta$	$\alpha+\beta$
proteins tested	1007	1219	1205	1369
with exceptions	0	26	8	9
total structures	8150	11467	11451	20984
exceptions	0	29	8	9

**Table 1.** Exceptions to our rules of metric consistency. Our definition of  $\beta$ -sheets tolerates exceptions, like bulges, which produces longer structures and more stringent demands. Even so, the exceptions are very rare.

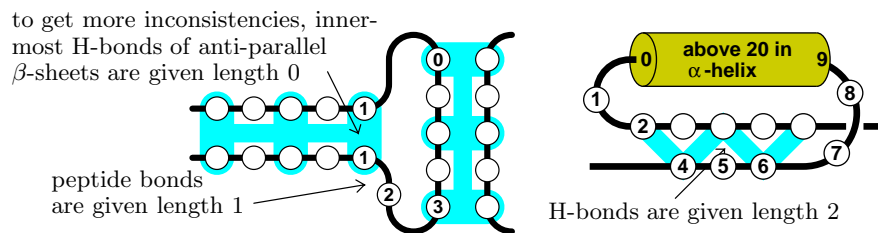
We define a pseudo-distance within a graph of protein residues, and by the rigidity of a secondary structure  $u$  we will mean the rigidity the subgraph  $H(u)$  within the protein graph  $G$ .

Our graphs have edges for peptide bonds, each with length 1, and for hydrogen bonds, each with length 2. In the latter case we make one exception: in an anti-parallel  $\beta$ -sheet, the hydrogen bond  $(i, j)$  with the minimal  $|i - j|$  has length 0. The idea behind that exception is to make it more difficult to consistently insert a strand of a structure inside the loop of an anti-parallel  $\beta$ -sheet. Note that without making this exception the vertical anti-parallel  $\beta$ -sheet in Fig. 5 would be rigid.

A set of structures is metrically consistent if their characteristic sets are disjoint and every one of them is rigid in the graph made by the peptide bonds and the hydrogen bonds of all of them.

It must be stressed that our pseudo-distance is somewhat proportional to the actual distances between different residues, but this proportionality is a subject of substantial discrepancies, especially inside  $\beta$ -sheets; the latter are often depicted as twisting ribbons. Nevertheless, the exceptions to the principle of metric consistency as we have stated are very few. To verify that, we have tested the proteins from PDB that were identified in ASTRAL files (see [5]) for the respective classes that have at most 30% aligned pairwise identity. The results of these tests are in Table 1.

In this section we will consider this question: does the problem of finding a maximum weight set of pairwise consistent secondary structures have a constant



**Fig. 5.** Examples of metric contradictions: inside the vertical  $\beta$ -sheet, the distance between top and bottom residues is 4, while outside there is a path of length 3; inside the  $\alpha$ -helix, the distance between the first and the last residue is 10 (or more), while outside there is a path of length 9.

factor approximation? To answer this problem in the positive, we will establish that the graphs that model this problem have a finite conflict number.

Because the structure of metric conflict can be complex, we first reduce this question to a simpler one. First, we define a *shell* for each structure; the shell of an  $\alpha$ -helix is its characteristic set, but for a  $\beta$ -sheet we make their intervals contiguous, *e.g.*, pair  $\{9, 11, 13, 15\} \cup \{24, 26, 28, 30, 32\}$  becomes  $\{9, 10, \dots, 15\} \cup \{24, 25, \dots, 32\}$ . Two secondary structures are *strictly non-overlapping* if their shells do not overlap.

Thus we can form three graphs for a set  $S$  of secondary structures: in  $G_0(S)$  the edges correspond to overlaps of characteristic sets, in  $G_1(S)$  the edges correspond to overlaps of characteristic sets and metric conflicts, and in  $G_2(S)$  the edges correspond to overlaps of shells and metric conflicts.

The conflict number of  $G_1(S)$  can be related to the conflict number in  $G_2(S)$  as follows:

**Lemma 3.** *For every independent set  $A$  in  $G_1(S)$  we can find a subset  $A'$  that is independent in  $G_2(S)$  such that  $|A'| \geq |A|/4$ .*

**Proof.** In the full version. □

**Lemma 4.** *For a structure  $u$  consider set  $\mathcal{N}(u)$  of structures that (a) have shells that do not overlap the shell of  $u$ , and (b) are in metric conflict with  $u$ . Then  $\mathcal{N}(u)$  does not contain a subset that is independent in  $G_2(S)$  with more than 4 elements.*

These two lemmas show that

**Theorem 2.** *The conflict number of  $G_1(S)$  is at most 24.*

**Proof.** Consider a structure  $u$  that has a smallest characteristic set  $I$ , and let  $J$  be the shell of  $u$ . We already know that at most 4 non-overlapping characteristic sets overlap  $I$ , and the argument also shows that at most 4 non-overlapping characteristic sets overlap  $J - I$  (note that the latter set is always smaller than the former).

Thus if  $N_1(u, S)$  contains an independent set  $A$  with  $k$  elements, at least  $k - 8$  of them do not overlap the shell of  $u$ , by Lemma 3 it contains a subset of at least  $k/4 - 2$  elements that have pairwise disjoint shells, and, by Lemma 4,  $k/4 - 2 \leq 4$ . □

The proof of Lemma 4 is in Appendix B. This lemma concludes the proof of Theorem 2 and yields the following

**Corollary 1.** *There exists an approximation algorithm with ratio at most 24 for finding a set of secondary structures that is pairwise metrically consistent and which runs in time  $O(n^2)$ .*

## 6 Total metric consistency

The notion of metric consistency can be useful in other ways as well.

Number one, metric consistency of a set of secondary structures can be easily checked. We may call this test *total metric consistency*.

Number two, one can investigate other computationally easy consistency tests that perhaps would be even more effective in eliminating wrong assemblies of structures.

Number three, quite a few secondary structures have very unambiguous predictions. Metric consistency allows to eliminate the tentative predictions that are inconsistent with the set of accepted predictions.

This approach can lead to a branch and bound algorithm which can be very practical if the unambiguous predictions reduce the problem size sufficiently.

The average number of secondary structures in a protein domain is 8 to 10 (except for  $\alpha + \beta$  class, where it is 15), thus even a few unambiguous predictions may decrease the searching space very drastically. For this reason one can use exact algorithms even if they do not run in polynomial time in the worst case.

## 7 Conclusions and open problems

We have established several ways in which prediction of secondary structures can be improved using global optimization. We expressed the goal of finding a consistent set of secondary structure predictions with the maximum likelihood as optimization problems that are NP-hard, but with constant approximation ratios.

We made some preliminary tests [11] that offer some insights about the possible further direction of this work. The tests were based on the work of Cheng and Baldi [6] who obtained predictions of  $\alpha$ -helices and  $\beta$ -strands as well as a “potential” for each pairing of  $\beta$ -strands, *i.e.*, for each  $\beta$ -sheet that can exist assuming that the previous predictions are correct. Then as a final stage they were optimizing the sum of potentials of predicted  $\beta$ -sheets using a greedy algorithm.

One observation that we could make is that the consistency problem that they considered was similar to one of our problems and it could be formulated as an IP program (with row generation methods). In most cases these programs were solved almost exactly and the quality of prediction was improved (2% increase of the correlation of predicted pairings with the true pairings). When we added the check for metric consistency, about 2% of the predictions were rejected which improved the correlation by ca. 0.2%. This indicates that it is far more important to find a good consistency requirement and the objective function, while the approximation results may be unimportant.

Our second observation is that a higher value of the objective function were not always indicating a better solution, as we have found a variant of greedy heuristic that lead to an almost twice larger improvement of the correlation while it was obtaining lower sums of potential than the original greedy heuristic.

There are three possible (and not exclusive) explanations. One is that the potential function is obtained on the basis of very fragmentary information about

the protein structure, so it may be not only unreliable in the sense of random noise, but it can be also biased.

Another reason is a potential function can correctly rank strand pairing in terms of “likelihood”, but it may incorrectly rank combinations of pairings. The reason can be explained with a little example. Suppose that we have 4 possible structures,  $a_1, a_2, a_3$  and  $a_4$ , and two maximal consistent sets,  $\{a_1, a_4\}$  and  $\{a_2, a_3\}$  (*e.g.*, we may have conflict pairs  $(a_1, a_2), (a_1, a_3)$  and  $(a_2, a_4)$ ). The following vectors of score values rank these structures in the same way:  $(5, 4, 3, 1)$  and  $(25, 16, 9, 1)$ . For the first vector we prefer solution  $\{a_2, a_3\}$ , as  $4 + 3 > 5 + 1$ ; for the second vector we prefer  $\{a_1, a_4\}$  as  $25 + 1 > 16 + 9$ . One can see that when we obtain “potential” from training a neural network (like in [6]) or through statistical analysis (*e.g.*, as used in [14]) it is much easier to obtain a good ranking of individual pairs than a good ranking of their groups.

The third possible reason is that the *folding process* (that creates the secondary structures) itself behaves like a greedy algorithm, with the contact established one by one, and each time the most likely contact is established. That the contacts are established sequentially was conjectured by Richardson [16], while recently Przytycka et al. [15] showed that Richardson’s conjecture is consistent with the data on experimentally known protein structure. Assuming that this conjecture is true, then the folding process can correspond to a greedy algorithm — but it is also possible that it corresponds to something like branch and bound algorithm, with the protein chain folding and unfolding until it finds a stable configuration. In particular, the improved criterion of greedy choice in [11] incorporated some assumptions about the folding process.

What may offer the easiest gain in the prediction quality is an improved set of consistency rules, as the experimentally known protein structures have many regularities that are very easy to check. The notion of metric consistency described in this paper is a step in this direction.

## 8 Acknowledgments

We thank Webb Miller for valuable comments and support and Arthur Lesk for inspiring discussions. Jieun Jeong was supported in part by NIH grant HG02238, Piotr Berman was supported in part by NSF grant CCR-0208821.

## References

1. R. Bar-Yehuda, M. Halldórsson, J. Naor, H. Shachnai, and I. Shapira. Scheduling split intervals. *Proceedings of SODA 2002* 732-741.
2. P. Berman and B. DasGupta. Multi-phase Algorithms for Throughput Maximization for Real-Time Scheduling. *Journal of Combinatorial Optimization* 4(3): 307-323, 2000.
3. P. Berman and M. Karpinski. On some tighter inapproximability results. *Proceedings of 26th International Colloquium on Automata, Languages and Programming*, Lecture Notes in Computer Science 1644(1999): 200–209.

4. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. Bourne. The Protein Data Bank. *Nucleic Acid Research* **28**:235-242 (2000).
5. J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL compendium in 2004. *Nucleic Acids Research* **32**:D189-D192, 2004.
6. J. Cheng and P. Baldi. Three-stage prediction of protein  $\beta$ -sheets by neural network, alignments and graph algorithms. *Bioinformatics* **21**:175-184, 2005.
7. H. M. Fooks, A. C. R. Martin, D. N. Woolfson, R. B. Sessions and E. G. Hutchinson, Amino Acid Pairing Preferences in Parallel Beta Sheets. *J. Molecular Biology* **356**:32-44, 2006.
8. J. Hästad. Clique is hard to approximate within  $n^{1-\epsilon}$ . *Acta Mathematica* **182**, 105-142, 1999.
9. T. J. P. Hubbard. Use of  $\beta$ -strand interaction pseudo-potentials in protein structure prediction and modeling. Proc. Biotechnology Computing Track, Protein Structure Prediction MiniTrack of 27th HICSS, IEEE Computer Society Press (1994):336-354.
10. T. J. P. Hubbard and J. Park. Fold recognition and ab initio structure predictions using hidden Markov models and  $\beta$ -strand pair potentials. *Proteins: Structure, function and genetics* **23**:398-402, 1995.
11. J. Jeong, P. Berman and T. Przytycka, Bringing folding pathways into strand pairing prediction, manuscript.
12. P. K. Mehta, J. Heringa, and P. Argos. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science* **4**:2517-2525, 1995.
13. J. Meiler and D. Baker. Coupled prediction of protein secondary and tertiary structure. *Proc. of the National Academy of Sciences of USA* **100**(2):12105-12110, Oct. 14, 2003.
14. M. Menke, E. Scanlon, J. King, L. Cowen, and B. Berger. Wrap-and-Pack: A New Paradigm for Beta Structural Motif Recognition with Application to Recognizing Beta Trefoils. *Journal of Computational Biology* **11**(6):777-795, 2005.
15. T. Przytycka, R. Srinivasan and G. D. Rose. Recursive domains in proteins. *Protein Science* **11**:409-417, 2002
16. J. S. Richardson.  $\beta$ -Sheet topology and the relatedness of proteins. *Nature* **268**:495-500, 1977.
17. B. Rost. PhD: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology* **266**:525-539, 1996.
18. A. I. Serdyukov. An algorithm with an estimate for the traveling salesman problem of the maximum. *Upravlyaemye Sistemy*, **25**:80-86, 1984, in Russian, English summary by Barvinok in G. Gutin and A.P. Punnen (eds.), *The traveling Salesman Problem and Its Variations*, Kluwer Academic Publishers (2002):591-594.
19. K. T. Simons, C. Kooperberg and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Molecular Biology* **268**:2-9-225, 1997.
20. H. Zhu and W. Braun. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of  $\beta$ -sheet formation in proteins. *Protein Science* **8**:326-342, 1999.

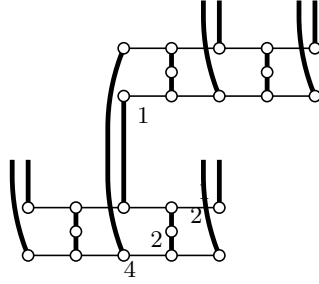


Fig. 6. Gadgets of two adjacent nodes.

## Appendix

### A 1-2 Matching is MAX-SNP hard

**Theorem 3.** *1-2 Matching is MAX-SNP hard, even if it is restricted to instances in which valid solutions cannot form cycles.*

Recall that an instance of 1-2 Matching is a graph in which each edge has a non-negative weight and a *width*, the latter being 1 or 2. A valid solution, or a 1-2 matching is a set  $M$  of edges such that no node is incident to edges of  $M$  with the sum of width exceeding 2.

A 1-2 matching forms a collection of node-disjoint paths and cycles, where paths (or cycles) that contain more than one edge consists only of edges of width 1. We will restrict our construction to instances in which edges of width 1 do not form cycles — and thus no valid solution may contain a cycle.

To show that 1-2 VWM is MAX-SNP hard, we will provide an approximation preserving reduction from MAX-CUT problem restricted to 3-regular graphs. An instance of MAX-CUT problem is an undirected graph, a valid solution is any node-set  $C$ , and the maximized objective function is the number of edges that have exactly one endpoint in  $C$  (and another in  $V - C$ ). MAX-CUT is MAX-SNP complete even when it is restricted to cubic graphs (see [3]).

Suppose that we have an instance  $G$  of MAX-CUT problem with  $2n$  nodes and  $3n$  edges, we will form an instance  $G'$  of 1-2 VWM with  $34n$  nodes and  $42n$  edges, and such that a cut in  $G$  with  $k$  edges will correspond to a 1-2 matching with weight  $32n + k$ . Moreover, given a 1-2 matching of weight  $w$  we will have a cut with at least  $w - 32n$  edges. Note also that  $G$  always has a cut with at least  $2n$  edges.

For a node  $u$  of  $G$  we construct a gadget, Fig. 6 shows a pair of such gadgets of two adjacent nodes. The horizontal edges in this diagram have width 1 and all other edges have width 2. A gadget  $\Gamma_u$  has a *selector node* that is on the very top (or the very bottom) of the gadget. The weights of edges on a particular level are all the same (4, 2 or 1).  $\Gamma_u$  consists of two parts, a black part with black *contact nodes* and a gray part with black-and-white contact nodes.

Finally, if nodes  $u$  and  $v$  are adjacent, we make two *connector edges* between the contact nodes, from a black contact  $\Gamma_u$  to a black-and-white contact of  $\Gamma_v$  and vice versa; connector edges are disjoint.

Suppose that in  $G$  there exists a cut  $C$  crossed by  $k$  edges. We form a 1-2 matching  $M(C)$  as follows: if  $u \in C$  then in  $\Gamma_u$  we vertical (or slanted) edges of the gray part

and horizontal edges of the black part; connector edges contained in  $M(C)$  can be adjacent to black contact nodes. If  $u \in V - C$ , we reverse the role of colors. When  $u \in C$  and  $v \in V - C$ , matching  $M(C)$  contains the connector edge from a black contact of  $\Gamma_u$  to a black-and-white contact of  $\Gamma_v$ .

Observe that vertical and horizontal edges in each part of the gadget have the same sum of weights, 8, so in every gadget we selected edges of total weight 32. Moreover, we selected exactly as many connector edges as there are edges in the cut of  $C$ .

Suppose that we have a 1-2 matching  $M$  in  $G'$ . We will *normalize* this matching so the weight does not decrease, and  $M = M(C)$  for some cut  $C$  in  $G$ .

- ❶ Consider the selector node of  $\Gamma_u$ ; if no adjacent edge is in  $M$  we can insert it, adding weight 4 (and removing at most 2 edges of joint weight at most 4).
- ❷ Consider a pair of horizontal edges that have weight 2. If adjacent edge of weight 4 is in  $M$  these edges cannot be in  $M$ , but otherwise we insert them, each insertion may require an edge removal, but no loss of weight is possible.
- ❸ Consider a pair of vertical edges that have weight 2. If adjacent edges of weight 2 are in  $M$  these edges cannot be in  $M$ , but otherwise we insert them, each insertion may require two edge removals, but only of edges of weight 1, so no loss of weight is possible.
- ❹ Consider a path of 4 horizontal edges that have weight 1. As in the previous cases, either we have already assured that none of them can belong to  $M$  or we can insert all of them to  $M$ .

## B Proof of Lemma 4

**Lemma 4** *For a structure  $u$  consider set  $\mathcal{N}(u)$  of structures that (a) have shells that do not overlap the shell of  $u$ , and (b) are in metric conflict with  $u$ . Then  $\mathcal{N}(u)$  does not contain a subset that is independent in  $G_2(S)$  with more than 4 elements.*

**Proof.** We will use  $A$  to denote interval(s) of the shell of  $u$ , and  $B$  to denote interval(s) of the shell of a structure that does not overlap  $A$ 's and is in a metric conflict with  $u$ . We will use  $X, Y, Z$  to denote intervals that separate  $A$ 's and  $B$ 's. We will also use lower case letters to denote the lengths of the respective intervals.

### B.1 $\alpha$ -helices

Two  $\alpha$ -helices that are not in conflict in  $G_0$  are not in conflict in  $G_2$ . Suppose that  $u$  under discussion is an  $\alpha$ -helix, then if it is in conflict with another structure  $w$ ,  $w$  must be a  $\beta$ -sheet, and their shells must have arrangement  $BXAYB$ . Suppose that  $w$  is an anti-parallel  $\beta$ -sheet; the metric inconsistency can be that  $A$  does not fit in the loop that is created by  $B$ 's. The condition for consistency is

$$x + y \geq a/2.$$

Suppose that  $w$  is a parallel  $\beta$ -sheet; then  $A$  cannot be too long nor too short, as we can traverse from one end of  $B$  to another by traversing  $X, A, Y$ , hence

$$x + a/2 + y \geq b,$$

and we can traverse from one end of  $A$  to another by going through  $Y, B, X$ , hence

$$x + y + b \geq a/2.$$

Because  $a \leq b$ , the latter is surely true.

Suppose that  $u$  is in conflict with two  $\beta$ -sheets, one with shorter strands  $B_0$  and with longer (or equal in length)  $B_1$ . Consider the arrangement of  $A$  and  $B_0$ ,  $B_0X_0AY_0B_1$ ; if  $x_0 \geq b_0$  or  $y_0 \geq b_0$  then  $A$  and  $B_0$  are consistent, hence  $B_1$ 's must be outside  $B_0$ 's. In that case, the only way  $B_1$  and  $A$  can be inconsistent is when  $B_1$  is parallel and  $x_1 + a/2 + y_1 < b_1$ .

However, both  $X$  and  $Y$  contain a strand  $B_0$ , so we can traverse from one end of  $B_1$  to another by traversing part of  $X$ , and then jumping to  $Y$  using a hydrogen bond of  $B_0$ , hence the consistency of  $B_0$  and  $B_1$  implies that  $x + y \geq b_1$ .

We conclude that in  $G_2(S)$  the independence number of  $\mathcal{N}$  is at most 1.

## B.2 Types of conflicts of $\beta$ -sheets

The previous subsection gave a simple example how metric inconsistency is obtained. Postulating two structures let us make a graph in which edges connect consecutive amino acids and the ‘‘pairs of parallel amino acids’’ from a postulated  $\beta$ -sheet (let us refer to these edges as bridges). In this graph we can find a path between two points in a postulated structure that is shorter than the shape of the structure allows.

Observe that we can restrict the bridges that we use to those at the ends of the respective  $\beta$ -sheets. Otherwise we can move the bridge in two directions and either we get a shorter path in one of the directions, or we have no difference in path length.

Observe also that if a path traverses a  $B$ , then this path cannot be a shortcut for a pair of ends of a  $B$  or an  $A$ . Therefore we never enter a  $B$ .

The first type of conflict we will call *strangling*: a path provides a shortcut between two end of an  $A$  without using a bridge of  $A$ . This implies a pattern  $BXAYB$  where  $B$ 's form an anti-parallel  $\beta$ -sheet and  $x + y < b$ . One can show that

- (a) there can be at most two consistent stranglings of one strand  $A$ ;
- (b) If a strand  $A$  is strangled  $i$  times, then it is overlapped by at most  $4 - 2i$  strands from a consistent set of structures (consistent in  $G_0$  sense).

[Two stranglings actually cannot co-exist and we can eliminate this possibility by introducing a rule that would create a conflict between two  $\beta$ -sheets that simultaneously strangle  $A$ . This rule would not change the correct estimate of the conflict number, but it can be useful in an application.]

We can conclude that if there exist strangling conflicts, we can decrease the estimate of the conflict number of  $u$  in  $G_1$ .

The second type of conflict we will call *internal*. The first subtype is *single-strand* conflict: we provide a shortcut between two ends of a  $B$  strand without using a  $B$ -bridge. This implies configuration  $AXBZA$  and either  $u$  is an anti-parallel  $\beta$ -sheet and  $x + z < b$ , or  $u$  is parallel and  $x + z + a < b$ . The second subtype is *two-stranded internal*, with pattern  $AXBYBZA$ . If  $u$  and  $w$  are both anti-parallel there is no conflict. If  $w$  is anti-parallel, then the shortcut can only connect ends of  $A$  so we have consistency if  $x + z < a$  (this assumes that  $u$  is parallel). If  $w$  is parallel, we cannot use bridges of  $B$  for a shortcut, so the conflict requires, if  $u$  is anti-parallel,  $x + z < b$ , and when  $u$  is parallel,  $x + z + a < b$ . Note that inequality for internal conflicts have the form  $x + z < c$  where  $c \in \{a, b - a, b\}$ .

Subsequently, we will assume that  $w$  does not have a single-stranded or constricting conflict with  $u$ .

The third type of conflict we will call *outer*, and it happens when the pattern is  $BXAYAZB$ . Again, there is no conflict when  $u$  and  $w$  are both anti-parallel, and

when  $u$  is parallel and  $w$  is anti the conflict requires  $x + z < a$ . When  $w$  is parallel, the conflict must have the form of a shortcut between the endpoints of  $B$ ; this requires  $x + z < b$  when  $u$  is anti-parallel and  $x + z < b - a$  when  $u$  is parallel.

The fourth type of conflict is *left straddle* which happens when the pattern is  $AXB_YAZB$ . A shortcut between the endpoints of  $A$  would require strangling, a shortcut between the endpoints of  $B$  that does not use  $Z$  would require a single-stranded conflict, a shortcut using  $Z$  and  $Y$  would exist even when we do not postulate  $u$  (hence  $w$  is in conflict with itself) so we need to address only the shortcut that uses both  $X$  and  $Z$ . When  $w$  is parallel, there is no conflict, so we assume it is anti-parallel. When  $u$  is anti, the conflict requires  $x + z < b - a$ , when  $u$  is parallel, it requires  $x + z < b$ .

The fifth and last type of conflict is *right straddle* and it requires the pattern  $BXAYBZA$ . It is strictly symmetric with the left straddle.

Since we do not count the cases of strangling conflicts, we have to count how many pairwise consistent conflicts we can have that belong to the other four types. We will show that for each type we may have only one.

### B.3 Internal conflict

We have arrangement  $AXB_0ZA$  or  $AXB_0YB_0ZA$  and the conflict is caused by a small value of  $x + z$  (the threshold being  $a, b_0 - a$  or  $b_0$ ).

One can see that  $B_1$  can be present neither in  $X$ , nor in  $Z$ . If  $w_0$  has a single stranded conflict, this does not allow  $w_1$  to be in inner or straddle type of conflicts, leaving only outer conflict as a possibility. If  $w_0$  has a two-stranded inner conflict, we must consider the case when  $B_1$  is present in  $Y$ .

Suppose that  $w_1$  also has an inner conflict. Then  $Y$  can be viewed as arrangement  $X'B_1Y'B_1Z'$ . One can see that if we have a shortcut between the ends of  $B_1$  that uses a bridge of  $A$ , there exists an even better shortcut that uses a bridge of  $B_0$ , hence  $w_0$  is in a metric conflict with  $w_1$ , a contradiction.

### B.4 Outer conflict

We have arrangement  $B_0XAYAZB_0$  and the conflict is caused by a small value of  $x + z$  (the threshold being  $a, b_0 - a$  or  $b_0$ ). Again,  $B_1$  can be present neither in  $X$ , nor in  $Z$ .

We can exclude the case when  $w_1$  also forms an outer conflict. Would it happen, we would have arrangement  $B_1X'B_0XAYAZB_0Z'B_1$ .

If  $w_1$  is anti-parallel, then conflict would require that  $u$  is parallel and  $x_1 + z_1 < a$ ; the latter is not possible because  $x_1 = x' + b_0 + x$ . Thus it remains to consider 3 cases:

Case:  $u$  anti-parallel,  $w_0$  parallel,  $w_1$  parallel.

$$x' + b_0 + x + z + b_0 + z' < b_1 \text{ (for } u \text{ and } w_1),$$

$$x' + z' \geq b_1 - b_0 \text{ (for } w_0 \text{ and } w_1).$$

The former inequality contradicts the latter

Case:  $u$  parallel,  $w_0$  parallel,  $w_1$  parallel.

$$x' + b_0 + x + z + b_0 + z' < b_1 - a \text{ (for } u \text{ and } w_1),$$

$$x' + z' \geq b_1 - b_0 \text{ (for } w_0 \text{ and } w_1).$$

The former inequality still contradicts the latter.

Case:  $u$  parallel,  $w_0$  anti-parallel,  $w_1$  parallel.

$$x' + b_0 + x + z + b_0 + z' < b_1 - a \text{ (for } u \text{ and } w_1),$$

$$x' + z' \geq b_1 \text{ (for } w_0 \text{ and } w_1).$$

The former inequality again contradicts the latter.

## B.5 Straddle conflict

Assume left straddle; we have arrangement  $AXB_0YAZB_0$  and the conflict is caused by a small value of  $x + z$  (the threshold being  $b_0 - a$  or  $b_0$ ). As before,  $B_1$  cannot be present in  $X$  or  $Z$ , so it can only be present in  $Y$  or outside. We also know that  $w_0$  is anti-parallel.

Suppose that  $w_1$  also forms a left straddle conflict. Then  $w_1$  is also anti-parallel and we have arrangement  $AXB_0Y'B_1Y''AZB_0Z'B_1$ . We have two cases:

Case:  $u$  anti-parallel,  $w_0$  anti-parallel,  $w_1$  anti-parallel.

$$\begin{aligned} x + b_0 + y' + z + b_0 + z' &< b_1 - b_0 \\ y' + z' &\geq b_1 - b_0 \end{aligned}$$

Case:  $u$  parallel,  $w_0$  anti-parallel,  $w_1$  anti-parallel.

$$\begin{aligned} x + b_0 + y' + z + b_0 + z' &< b_1 \\ y' + z' &\geq b_1 - b_0 \end{aligned}$$

In either case, the inequalities contradict each other.  $\square$

## C Lemma 3

**Lemma 3.** *For every independent set  $A$  in  $G_1(S)$  we can find a subset  $A'$  that is independent in  $G_2(S)$  such that  $|A'| \geq |A|/4$ .*

**Proof.** We can form set  $A'$  as follows: start with  $A' = \emptyset$ ; find  $u \in A$  with the least number of neighbors in  $G_2(S)$ ; insert  $u$  to  $A'$ , remove  $u$  and its  $G_2(S)$ -neighbors from  $A$ ; continue while  $A \neq \emptyset$ .

One can see that the claim holds if in each iteration we remove at most 4 elements from  $A$ , *i.e.*, if in each iteration we find  $u \in A$  with at most 3 neighbors. We can show a contradiction if  $A \neq \emptyset$  and yet no  $u \in A$  has fewer than 4 neighbors. Because  $A$  is metrically consistent and its elements have disjoint characteristic sets, the only reason for  $u \in A$  to have neighbors is that the segments of its shell may overlap with other shell segments.

Define a relation:  $u \rightarrow u'$  means that a segment of the shell of  $u$  contains an entire segment of the shell of  $u'$  and that the latter shell has fewer elements. Clearly, relation  $\rightarrow$  defines chains of limited length only.

Observation: if a segment of the shell of  $u$  intersects three other shells, then for some  $u'$  we have  $u \rightarrow u'$ . Indeed, let  $I = \{i, \dots, j\}$ , numbers  $i, i+2, j-2, j$  belong to the characteristic set of  $u$ , numbers  $i+1, j-1$  belong to some two shells, so the third intersecting shell segment, say of  $u'$ , must be contained in  $\{i+3, \dots, j-2\}$ ; as it has at most  $j-i-5$  elements, the other segment of the shell of  $u'$  has at most  $j-i-3$  elements and thus the shell of  $u'$  has at most  $2(i-j)-5-3$  elements, while the shell of  $u$  has at least  $2(i-j)$  elements.

Observation: if  $u \rightarrow u'$  then for some  $u''$  we have  $u' \rightarrow u''$ . Indeed, the segment of  $u'$  that is included in a segment of  $u$  intersects only one shell of a structure from  $A$ ; hence the other segment of  $u'$  intersects at last three shells, so we can apply the previous observation.

Observation: no shell segment intersects more than two other shells. Indeed, the previous two observations show that this would imply an infinite chain of  $\rightarrow$  and thus a contradiction.

Now consider a segment  $\{i, \dots, j\}$  of a shell  $u$  from  $A$  with the smallest  $j$ ; this segment can intersect at most one other shell, and since the other segment of the shell of  $u$  intersects at most two shells, the shell of  $u$  intersects at most 3 other shells.  $\square$