Construction of Orthogonal CC-sets

Andrej Brodnik, Marko Palangetić and Daniel Silađi
University of Primorska, Koper, Slovenia
E-mail: andrej.brodnik@upr.si, palangeticmarko95@hotmail.com, szilagyi.d@gmail.com

Vladan Jovičić
École Normale Supérieure, Lyon, France
E-mail: vladan94.jovicic@gmail.com

Keywords: algorithms, NP-hardness, modeling

Received: October 30, 2018

In this paper we present a graph-theoretical method for computing the maximum orthogonal subset of a set of coiled-coil peptides. In chemistry, an orthogonal set of peptides is defined as a set of pairs of peptides, where the paired peptides interact only mutually and not with any other peptide from any other pair. The main method used is a reduction to the maximum independent set problem. Then we use a relatively well-known maximum independent set solving algorithm which turned out to be the best suited for our problem. We obtained an orthogonal set consisting of 29 peptides (homodimeric and heterodimeric) from initial 5-heptade set. If we allow only heterodimeric interactions we obtain an orthogonal set of 26 peptides.

Povzetek: V članku je predstavljen izračun največje ortogonalne množice peptidov z uporabo metod teorije grafov. Za elemente ortogonalne množice velja, da, če dva elementa vzajemno delujeta, potem ne delujeta z nobenim drugim elementom množice. Algoritem, ki uporablja prevedbo na problem največje neodvisne množice, je bil uporabljena v praksi.

1 Motivation

In the last 30 years, impressive 3D structures have been built using DNA, in a field called DNA origami [5]. Complex structures built from proteins would have many advantages, since amino acids provide much more functionality. The main problem is that the simple Watson-Crick paring rules present in DNA have no simple analogue for proteins. Using a special class of polypeptides, called coiled-coil polypeptides, the orthogonal binding rules of DNA can be emulated [2, 4]. By specifying only the primary structure of those polypeptides (the order of amino acids), complex 3D structures can be built, such as the recent protein tetrahedron [3]. More specifically, that structure is determined by taking the wireframe of the desired object, doubling every edge, and performing an Euler traversal of the obtained graph. Then, we know that the peptides associated with edges that were initially parallel must bind, and all others must not.

Essential for such designs is that each pair of peptides interacts only mutually, and not with any other pair. Thus, the notion of an *orthogonal set* is introduced. Obviously, the greater our orthogonal set is, the more complex are the structures we can create. Currently the limiting factor in designing larger structures is the small set of available peptides.

In this paper, we describe a method for determining an orthogonal set of maximum size, from a given set of admissible peptides. Also, in section 6 we present a possible approach for extending a given orthogonal set.

2 Problem description

As input we are given a set of peptides $P = p_1, p_2, \dots p_n$ (their primary structures – given as strings of fixed length) and interaction matrix I. If $I_{i,j} = 1$, then p_i interacts with p_j and if it is 0 they do not interact. We have to construct a set of pairs S, where $(p_i, p_j) \in S$, iff $I_{i,j} = 1$ and for all other p_k that are in any pair of S $I_{i,k} = 0$. Moreover, if $i = j$ in (p_i, p_j) we are talking of *homodimer* and otherwise of *heterodimer*.

We can model this problem as a graph-theoretical one: First, an undirected graph $G = (V, E)$ where V is the set of peptides P, and the edge set E contains an edge $p_i p_j$ (or a loop at p_i, denoted by $p_i p_i$) if and only if p_i and p_j interact. Given that graph, we want to find a subset of non-adjacent edges whose vertices are also non-adjacent. More formally, and more conveniently for later consideration, our problem can be defined as follows:

Definition (Maximum Independent Set of Pairs (MISP)). Let $G = (V, E)$ be an undirected graph and let k be a positive integer. Does there exist set a $S \subseteq E$ such that for any $u_1 v_1, u_2 v_2 \in S$

$$\{u_1, v_1\} \cap \{u_2, v_2\} = \emptyset,$$

$$\{\{u_1, u_2\}, \{u_1, v_2\}, \{v_1, u_2\}, \{v_1, v_2\}\} \cap E = \emptyset$$

and $|S| > k$?

3 Hardness of the problem

In order to determine the best possible solution of our problem, in this section we will prove that MISP is NP-complete.

Theorem 1. *Maximum independent set of pairs is NP-complete.*

Proof.

Algorithm 1 NP certifier

```
$S \leftarrow$ given set of pairs
if $|S| < k$ then
    return No
for $u_1v_1 \in S$ do
    for $u_2v_2 \in S - u_1v_1$ do
        if $u_1u_2 \in E \vee u_1v_2 \in E \vee v_1u_2 \in E \vee v_1v_2 \in E \vee u_1v_1 \notin E$ then
            return No
return Yes
```

The problem is easily seen to be in NP, and Algorithm 1 is its polynomial certifier.

Now we will reduce the *independent set* problem to MISP in order to show that MISP is NP-hard.

Let $G = (V, E)$ be a graph, for which we want to check if there exists an independent set of size greater than k. Define a new graph $G' = (V', E')$ as follows. Initially, let $V' = V$ and $E' = E$. Then, for each vertex $v \in V$ add another vertex v' (a copy of v) to V' and add the edge vv' to E'.

Lemma 1. *Every maximal independent set of pairs in G' consists only of the edges of the form vv'.*

Proof. Let S be a MISP in G'. Suppose the contrary, i.e., there is a pair $uv \in S$ which is not of the form ww' for $w \in V$. Then, for all $u_1v_1 \in S$ we have $u_1u \notin E'$, $u_1v \notin E'$, $v_1u \notin E'$, $v_1v \notin E'$. Then we can delete the pair uv from S and add pairs uu' and vv' where u' and v' are copies of u and v, respectively. We can do this since the only neighbors of u' and v' are u and v, respectively. We obtained an independent set of pairs, with more more than $|S|$ elements, a contradiction. □

Now we will prove that there is an independent set $|S| \geq k$ in G if and only if there is an independent set of pairs $|S_P| \geq k$ in G'.

($\Rightarrow$) Suppose that S is an independent set in G with $|S| \geq k$. Then, define the independent set of pairs S_P in G' on the following way:

$$S_P = \{vv' \mid v \in S\}.$$

It is easy to verify that this is independent set of pairs according to the definition above. Then $|S_P| = |S| \geq k$.

($\Leftarrow$) For the other direction, suppose that S_P is an independent set of pairs in G' with $|S_P| \geq k$. Then, by the previous lemma, we can define the following independent set S in G:

$$S = \{v \in V \mid vv' \in S_P\}.$$

By the construction of graph G' and by the lemma, one can show that S is an independent set of G. Then $|S| = |S_P| \geq k$ which completes proof that MISP is NP-hard.

Combining the NP-hardness with the earlier fact that MISP is in NP, we conclude that MISP is NP-complete. □

4 Reducing MISP to the maximum independent set

Now that we know that MISP is NP-complete, we can use one of the vast number of algorithms already developed for solving various NP-hard problems, once we reduce MISP to that problem. The most natural choice is the maximum independent set problem [1, 6, 7].

Based on the MISP graph $G = (V, E)$, we construct a new graph $G^* = (V^*, E^*)$, where $V^* = E$, and two vertices are connected (in G^*) if and only if their corresponding edges in G share a common vertex or have two of their vertices connected by an edge. It is easy to see that finding an independent set in G^* will give us an independent set of pairs, according to the definition in section 2. Moreover, due to our construction, an independent set of pairs in G also gives us an unique independent set in G^*.

Thus, we have obtained a bijection between the independent sets of G^* and the independent sets of pairs of G.

5 Results

We use results from the previous section to solve the MISP of the input graph G which is constructed from the input set of peptides $P = p1, p_2, \ldots, p_n$ in several steps.

1. Based on previous work by [4], we calculate the interaction scores s_{ij} for each pair of peptides p_ip_j (including homodimers p_ip_i), and store that matrix for the following steps.

2. Choose a threshold t based on which we decide whether peptides p_i and p_j with interaction score s_{ij} will interact. If $s_{ij} < t$, we declare that p_i and p_j are not interacting (or, more precisely, interacting in a negligibly small proportion), and likewise, if $s_{ij} \geq t$, p_i and p_j interact. For practical purposes, we might want to introduce two thresholds t_s (strong interaction threshold) and t_w (weak interaction threshold) such that $t_s - t_w$ is a positive "safety margin" accounting for the inexactness of the scoring function from the previous step. Then, the vertices of G^* would be just the pairs that interact strongly, but they will be connected even if they interact weakly. However, such considerations are outside the scope of this paper.

3. Construct the graph G on the set of peptides by connecting the interacting ones, as described in section 2.

4. Reduce G to G^*, suitable for computing the independent set, as described in section 4.

5. Find the maximum independent set in G^*. As shown before, it corresponds to the MISP (or, orthogonal set) in G. We use the (exact) maximum clique solving algorithm presented in [1], which is based on greedy graph colorings – i.e. if we can color a particular subgraph with k colors, we know that that the maximum clique in that subgraph has size at most k.

In order to test our algorithm, we generated synthetic initial sets of peptides, based on two observations: Firstly, the interaction scoring function is designed to consider only 4 positions in each heptad. Secondly, using electrostatic arguments about individual amino acids and their positions in the coiled-coil, we reduced the variation even further, by allowing only 2 different amino acids on 3 of those 4 positions, and completely fixing the remaining amino acid. Thus, we obtain 8 essentially different heptads, which we use to build up larger peptides. Our main result is the calculation of a 29-peptide orthogonal subset of the 5-heptad initial set (2^{15}) peptides (generated as described above), as well as a 26-peptide purely heterodimeric orthogonal subset of the same initial set. The interaction score heatmap can be seen on Figures 1 and 2.

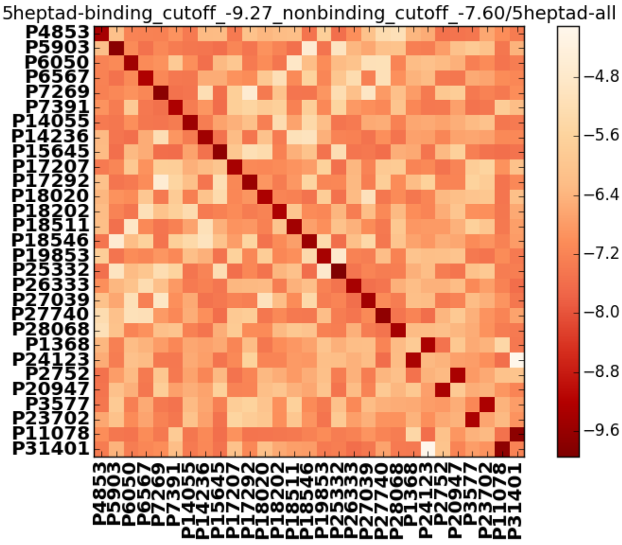

Figure 1: 5-heptad orthogonal set, no restriction

The peptidets which belong to orthogonal set are in both figures colored in dark red.

6 Future work

Up to now, we have only considered orthogonal sets derived from synthetically generated peptides, as described in the previous section. To actually use such an orthogonal set, we had to manually synthesize all of those peptides.

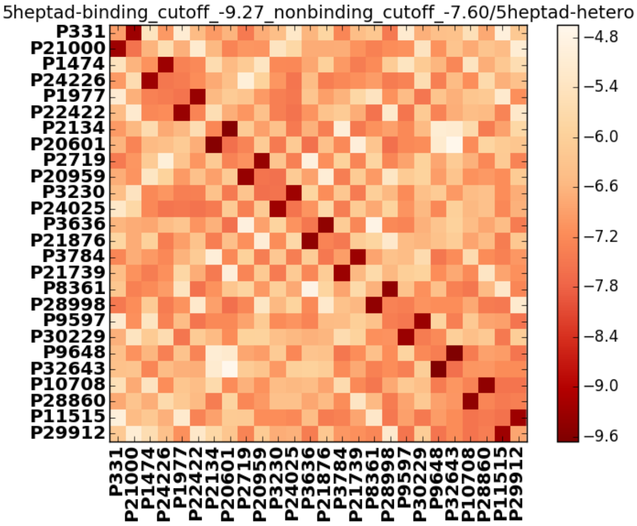

Figure 2: 5-heptad orthogonal set, heterodimers only

An alternative would be to construct a maximal orthogonal set from the set of all natural tetraheptads (coiled-coils where each of the 4 heptads occurs naturally). Since there are 1171 known natural heptads, we can combine them to get $1171^4 = 1\,880\,301\,880\,081$ possible tetraheptads. Finding a maximal orthogonal subset of this set would require finding the maximum independent set of a graph with more than 10^{12} vertices – a task clearly impossible to do in a reasonable amount of time.

Either way, we generate a large number of "useless" peptides, that will be discarded later, and not used in the (much smaller) orthogonal set.

Our idea is to use a heuristic to reduce the initial set to a more manageable size: since it is possible to calculate the interaction matrix for single natural heptads, we can approximate scores for tetraheptads as shown at Figure 3. More specifically, we will add up the precalculated scores between (adjacent) heptads which are connected as on figure 3. Of course, some interactions will be left unaccounted for in the final score, for example the last amino acid in heptad 1 on 3 may interact with first amino acid of heptad 7 which is not added to the final score.

This observation enables us to construct more meaningful initial peptide sets consisting of longer peptides, based on the already-calculated orthogonal sets of shorter peptides.

7 Concluding remarks

In this paper, we presented an exact method for determining an orthogonal set of coiled-coil polypeptides, if we are given a numeric measure of their interaction strength. Our approach has been demonstrated to be successful for moderately large initial peptide sets (tens of thousands), and has

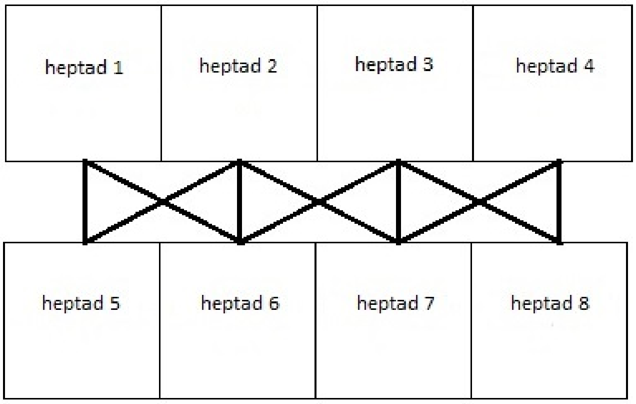

Figure 3: Proposed way of scoring

given us optimal orthogonal sets that could not have been calculated by hand.

Unfortunately, for even larger initial sets, the maximum-clique solver becomes an apparent bottleneck, as it has to operate on graphs of size $O(n^4)$, where n is the size of the initial set. In that case, we suggest investigating a bottom-up method described in the section 6.

References

[1] M. Depolli, J. Konc, K. Rozman, R. Trobec, and D. Janezic. Exact parallel maximum clique algorithm for general and protein graphs. *Journal of chemical information and modeling*, 53(9):2217–2228, 2013. https://doi.org/10.1021/ci4002525.

[2] J. H. Fong, A. E. Keating, and M. Singh. Predicting specificity in bZIP coiled-coil protein interactions. *Genome biology*, 5(2):R11, 2004. https://doi.org/10.1186/gb-2004-5-2-r11.

[3] H. Gradišar, S. Božič, T. Doles, D. Vengust, I. Hafner-Bratkovič, A. Mertelj, B. Webb, A. Šali, S. Klavžar, and R. Jerala. Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nature chemical biology*, 9(6):362–366, 2013. https://doi.org/10.1038/nchembio.1248.

[4] V. Potapov, J. B. Kaplan, and A. E. Keating. Data-driven prediction and design of bzip coiled-coil interactions. *PLoS Comput Biol*, 11(2):1–28, 02 2015. https://doi.org/10.1371/journal.pcbi.1004046.

[5] P. W. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006. https://doi.org/10.1038/nature04586.

[6] P. San Segundo, F. Matia, D. Rodriguez-Losada, and M. Hernando. An improved bit parallel exact maximum clique algorithm. *Optimization Letters*, pages 1–13, 2013. https://doi.org/10.1007/s11590-011-0431-y.

[7] E. Tomita, Y. Sutani, T. Higashi, S. Takahashi, and M. Wakatsuki. A simple and faster branch-and-bound algorithm for finding a maximum clique. In *International Workshop on Algorithms and Computation*, pages 191–203. Springer, 2010. https://doi.org/10.1007/978-3-642-11440-3_18.