A Similar Fragments Merging Approach to Learn Automata on Proteins

François Coste, Goulven Kerbellec

A Similar Fragments Merging Approach to Learn Automata on Proteins

François Coste and Goulven Kerbellec*

Symbiose, IRISA, Campus de Beaulieu,
35042 Rennes Cedex, France
{Francois.Coste, Goulven.Kerbellec}@irisa.fr

Abstract. We propose here to learn automata for the characterization of proteins families to overcome the limitations of the position-specific characterizations classically used in Pattern Discovery. We introduce a new heuristic approach learning non-deterministic automata based on selection and ordering of significantly similar fragments to be merged and on physico-chemical properties identification. Quality of the characterization of the major intrinsic protein (MIP) family is assessed by leave-one-out cross-validation for a large range of models specificity.

1 Introduction

Proteins are essential to the structure and function of all living cells and viruses. They are amino acid chains that fold into three-dimensional structures. Most of the times, only the amino acid chain – a sequence over 20 letters each representing one amino acid – is known. Determination of the structure or the function of proteins from their sequences is one of the major challenges in molecular biology. One of the most successful approaches is to define signatures of known *families* of biologically related proteins (typically at the functional or structural level). A representative example of this approach is the well-known Prosite database [1] gathering patterns defined essentially by experts for a large number of protein families. Automatic Pattern Discovery is a dynamic research field [2, 3]. Among the state-of-the-art algorithms, Pratt [4] (chosen to be the default pattern discovery tool proposed on the Prosite web site), Teiresias[5] or Splash[6] have been successfully designed to generate Prosite patterns, i.e. sub-regular expression, while, concerning stochastic models, the corresponding state of the art would be training profile hidden Markov models (which are left-right hidden Markov models focusing on so-called "match" positions and handling deletions or insertions of symbols) as in the commonly used tools HMMER [7] and SAM [8]. An important feature of these approaches is that they are limited to *position-specific* characterizations: neither relations between positions – for instance, if we consider the disulfide bond between cysteines, the fact that when a cysteine amino acid is present at position i there should be necessarily another cysteine at position j – nor alternative paths (disjunction over more than one position) can be

* Goulven Kerbellec is supported by a PhD research grant from Région Bretagne.

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 522–529, 2005.

represented, whereas it could be done in true regular expressions or automata. We address in this paper the task of learning *automata* for the characterization of proteins families to overcome the current position-specific limitation of Pattern Discovery. Learning automata has been widely studied in Grammatical Inference, notably by state merging techniques whose more representative algorithm is certainly RPNI [9, 10]. RPNI has been shown to have identification properties and good performances on artificial data. The number of needed data may be reduced with the help of the EDSM heuristic which has won the Abbadingo competition, still on artificial data. In contrast, while the application to genomic sequences seems to be a promising field for Grammatical Inference, not much work has been published on this matter (if we restrict ourselves to methods actually discovering a grammar and not just training its weight parameters, which would for instance exclude the work of Sakakibara on stochastic context free grammars for the prediction of RNA structure [11] but would include the application of Sequitur [12] to infer a hierarchical structure on DNA sequences without generalization capabilities). Concerning the application of such methods for the characterization of proteins, we are only aware of the early work of Yokomori [13] on learning locally testable languages, a subclass of automata which may be linked to *n-grams* and to persistent splicing systems, for the identification of protein α-chain regions.

Our main contribution in this article is the proposition of a new heuristic approach in the state-merging framework allowing a successful inference of automata for the characterization of proteins. The approach, sketched in Algorithm 1, consists of two main stages: first a *characterization* stage, introduced in section 2, detects and orders similar protein fragment pairs, then a *generalization* stage, described in section 3, merges the candidate fragment pairs to identify globally conserved areas and physico-chemical properties. We present a first validation of our approach on a real task of protein characterization in section 4. Technical details are omitted here due to space limitation. We refer interested readers to the associated technical report [14] for more details.

Algorithm 1. Significantly Similar Fragment Merging Approach

procedure SFP_MERGING(S: set of sequences, q: quorum, $\mathcal{G}$: set of amino acid groups, λ_G,λ_Σ: likelihood tests thresholds)

▷ Characterization stage (section 2)

$L \leftarrow$ LIST_OF_SFP(S)

L.SORT_BY_REPRESENTATIVITY_SCORE(S)

▷ Generalization stage (section 3)

$A \leftarrow$ MAXIMUM_CANONICAL_AUTOMATA(S)

for each *sfp* $\in L$ **do** ▷ Merging Fragment Pairs

  A.MERGE_IF_ADMISSIBLE(*sfp*)

A.GAP_GENERALIZATION(q) ▷ Representative Fragments

A.INFORMATIVE_POSITIONS($\mathcal{G}, \lambda_G, \lambda_\Sigma$) ▷ Physico-chemical Properties

return A

2 Characterization

Significantly similar fragment pairs. Our method relies on a set of significantly similar fragment pairs (SFP) for the characterization stage. When considering protein sequences, such a set can be extracted from the sequences by DIALIGN2 [15]. DIALIGN2 is a multiple alignment tool whose first step consists in finding all fragment pairs such that their similarity is significantly larger than expected on random sequences. In DIALIGN2, these SFP are then combined to make a multiple alignment optimizing the global sum of weights under consistency constraints. In our approach, this set of SFP is considered as a first selection of interesting fragments such that merging them is potentially interesting.

Ordering Similar Fragment Pairs. The selection of these fragments is based only on sequence-to-sequence comparison. We introduce two scores, detailed in [14], to rank these fragment pairs according to their representativeness of the whole protein family. The first score estimates the support in other sequences of the family, i.e. it counts for each SFP the number of sequences containing a fragment sufficiently similar to it. The second score relies on a set of proteins not belonging to the family to give priority to discriminative characteristic SFP [16]. It evaluates how the support of the SFP in the family implies its proportion to be supported in the family and in the other set of sequences. Each score defines an heuristic ordering of the SFP. We will refer to these ordering as being respectively the *support heuristic* and the *implication heuristic*.

3 Generalization

Merging Fragment Pairs. The first generalization step applies the classical state-merging scheme popularized by RPNI [9] and EDSM [17] to SFP. We consider the more general case allowing to learn non-deterministic automata. Following the definitions of [18], to which we refer the reader for details, the general sketch of this kind of algorithm is to first construct an automaton, named *maximal canonical automaton* (MCA), representing exactly the training set of sequences and, then, to generalize the recognized language by *merging* (unifying) some of its states. State merging algorithms can be distinguished by their choice of states to merge. We propose here to merge iteratively the states corresponding to the SFP identified in the characterization stage, beginning by SFP with higher representativeness. This ordering is taken into account by introducing a preservation constraint over the previously merged fragments. Namely, after each SFP merging, a constraint stating that the resulting states can not be merged together is set. Further SFP mergings that would violate such constraint are discarded.

Representative Fragments. Merging the SFP allows to identify hot spots: sets of contiguous positions where lots of fragments have been merged. Besides, some positions may be involved in none of SFP merges. These latter localizations are clearly not representative of the family. We propose to treat them as "gaps". We

introduce classically a quorum parameter. If a state is used by less sequences than specified by the quorum, it is merged with its neighbors. This step allows to keep only the characteristic regions and is an important generalization step for long proteins. Several variations around this scheme could be implemented. Statistical information like the length or the amino acid composition of the gap could also be considered and added to the model. The version presented here is the simple one used in the experiments.

Identification of Physico-chemical Properties. We propose here to recover the important physico-chemical properties of the amino acids at each position of the representative fragments with respect to the function or the structure of the family. The approach takes as input a set $\mathcal{G}$ of eventually overlapping substitution groups representing important physico-chemical properties (typically the groups proposed by Taylor [19]). For each position, likelihood tests are used to decide whether the set of amino-acids should be expanded to the smallest group including the set or else whether it should be expanded to the whole alphabet Σ if the distribution of the amino-acids appears to be random (see [14]). These tests introduce two threshold parameters λ_G and λ_Σ allowing to tune the risk when expanding to a group or else to Σ.

4 Experiments

We evaluated our approach on the major intrinsic protein (MIP) family [20]. The MIP family has functional and structural properties such as transmembrane channels, well-known to be important for water, alcohol and small molecules transport across cell membranes thanks to P. Agre (Nobel Prize in Chemistry "for the discovery of water channels", 2003). UNIPROT, a biological protein sequence database, contains 911 proteins annotated as being members of the MIP family. Of these 911, 159 protein sequences (denoted hereafter by the set T) are present in SWISSPROT which is the reliable annotated public reference database used by Prosite. Of this set, a biologist expert has identified only 79 sequences with a real biological experiment-based annotation (a lot of proteins being annotated "by similarity"). By filtering out the sequences with more than 90% of identity, this set was then reduced to 44 sequences (set M). Of this set, the expert has identified 24 water-specific sequences (set W+) and 16 glycerol or small molecule facilitator sequences (set W-). Let us notice the difficulty of the discrimination task between these MIP, some sequences of W+ being closer to some sequences of W- than to the other sequences of W+. We have established also a control set composed of sequences close to MIP sequences and identified by the expert as being outside the family (set C).

All the experiments were performed with an implementation of our approach named Protomata-L using DIALIGN2 with the following options : `-nta -thr 5 -afc`. The group expansion of Protomata-L has been done with the sets of physico-chemical properties proposed in Fig. 5 of [19] except the "unions" group, and $\lambda_G = 10^{-7}$, $\lambda_\Sigma = 10^{-19}$. Even with our unoptimized code, the execution never exceeded 10 minutes on a 3GHz desktop station.

Table 1. Comparison of 4 MIP signature patterns

Method	Precision	Recall	F-mes.	Pattern
Prosite (reference)	95%	91%	0.93	[HNQA]DNP[STA][LIVMF][ST][LIVMF][GSTAFY]
Pratt	90%	78%	0.83	GX(2)[FILMV]NP[AS]X[DST][FIL][AGP]
Teiresias	23%	89%	0.37	[ILMV]X(10)[ST]X(3)[ILMV]NX[AG]X(3)[AG]
Protomata-L	100%	87%	0.93	[ACGSTV]X[ACFGILMV]N[ACGPV][AGS][ACFG ILMV][DNST][ACFGILMV][ACGSTV]X[ACFGHI KLMTVWY]X(12)[FMY]X[ACFGHIKLMTVWY] XQ[ACFGHIKLMTVWY][ACFGILMV][AGS][AGS]

First Common Fragment. For this first set of experiments, in order to compare our fragment merging approach with Pratt[4] and Teiresias[5] methods and Prosite hand-made pattern[1], we restricted Protomata-L to return only the first common fragment shared by all sequences, using support heuristic. Pratt and Teiresias were used with their default parameters, except the parameter W (maximum length) of Teiresias that was set to 50 to allow longer pattern to be discovered. The patterns were learned from the set M and tested on the set T. Even if the set T was used in Prosite for the design of the pattern, a scan of the Prosite's pattern on SWISSPROT database returns false positive as well as false negative sequences with respect to T. Table 1 summarizes the results of such scans for the three patterns. The recall of our approach is close to Prosite's pattern recall while our precision remains at 100%. Let us remark that in our false positives, one was not a full sequence and 16 were annoted as MIP by similarity. When comparing our approach with Pratt and Teiresias, the comparison is clearly in favor of Protomata-L with respect to both the precision and the recall.

Water-Specific MIP subfamily. In this second set of experimentations, we focused on the characterization of the water-specific MIP subfamily set W+, using the set W- as counter-example. This discrimination task is motivated by a better understanding of the transport of these molecules. We used it to study the quality of the characterization on closely related sets of sequences at increasing specificity levels. Due to the small number of available sequences, a leave-one-out cross-validation scheme was used to evaluate our approach. For each couple of positive and negative sequences $(w+, w-)$, the training was achieved using the remaining sequences of W+ and W-. For each leave-one-out datasets, several automata – ranging from short automata (like in the previous paragraph) to larger automata characterizing almost all the length of the MIP topology – were obtained by using an increasing number of SFP. Each automaton was then evaluated according to the distance for acceptation of the positive sequence left out $w+$, the negative sequence left out $w-$, and also of the closest sequence c in the control set C. The *distance for acceptation* refers here to the minimal cost of amino acid substitutions needed in the sequence for its acceptation by

[1] Preliminary tests, not reported here, showed that RPNI and EDSM were not able to propose pertinent automata from this kind of data.

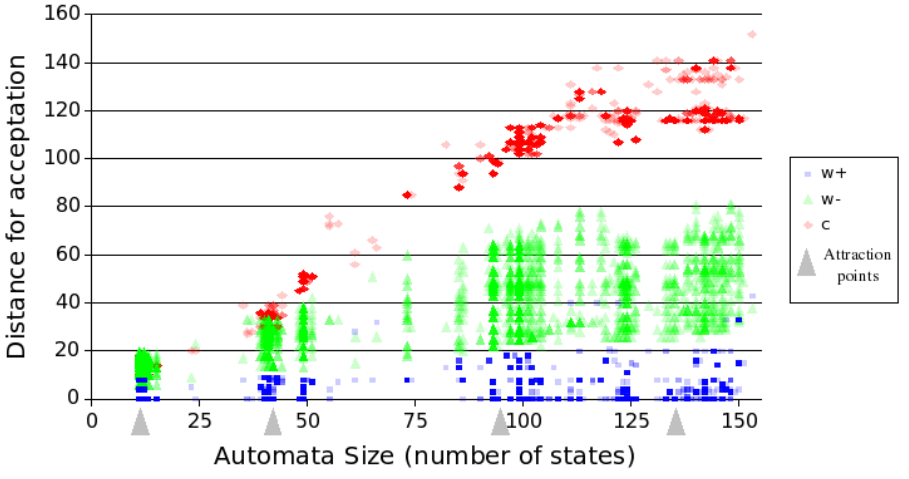

Fig. 1. Characterization of the Water Specific MIP family

the automaton (the cost of each amino acid substitution being given by the classical substitution matrix Blosum62 [21]). Fig. 1 presents the results of all these experiments when using the implication heuristic and a quorum of 100%. On the size axis, we highlighted 4 attraction points which are related to the progressive emergence of common sub-patterns, the first one corresponding to the first common fragment. The separation of the different sets of sequences is manifest and grows along the automata size axis until an inflexion point near 100 states. Behind this inflexion point, the merged SFP do not contribute anymore to the discrimination but only to a more precise characterization of the family without showing over-generalization evidence. Table 2 sums up the results of the automata at the attraction points for the classification task between W+ and W-, with strict acceptation and with a distance threshold acceptation. In the latter case, the closest counter-example distance from the automata was taken as the threshold distance for acceptation. The approach was then able to raise 100% of precision and 100% of recall for automata sizes of 40 or even 100 states.

Table 2. Performance on classification task (W+ vs W-)

	Strict			Threshold		
Automata Size	Precision	Recall	F-mes.	Precision	Recall	F-mes.
10	100%	92%	0.96	100%	96%	0.98
40	100%	71%	0.83	100%	100%	1.00
100	100%	54%	0.70	100%	100%	1.00
130	100%	42%	0.59	100%	96%	0.98

5 Conclusion

This study shows – even if it has to be confirmed on other sets of sequences – that good automata can be learned successfully on proteins. The proposed heuristic approach can be applied to the characterization of a family of proteins from positive examples only. It is also able to benefit from available counter-examples to produce more subtle models performing well in the discrimination of a closely related family of sequences. Depending on the application, the level of precision of the learned models can be chosen, ranging from short characteristic models (for classification tasks) to more detailed and explanatory models (for modeling the family of sequences). As proved by performance in leave-one-out cross-validation, the more specific models have still good prediction accuracy when allowing a small distance for acceptation to compensate the limited number of available examples. An alternative way to handle unpredictable family variation would be to use the learned automata as the underlying structure of probabilistic automata, or hidden Markov models, and estimate their stochastic parameters by the classical well-studied training methods. The advantage of our approach is that these variations are treated outside the model by measuring the distance to it, allowing the models to focus only on an explicit characterization of the important properties of the training sequences. We think that we could even improve the prediction accuracy by using distances taking into account the weights of the amino acids at each position with respect to the training sequences, but this has still to be implemented and tested.

Compared to classical protein Pattern Discovery algorithms, our approach introduces several new ideas. Globally, we think that, besides the ability to learn a more expressive class of model, the fundamental difference of Protomata-L with these Pattern Discovery approaches consists in the introduction of the similarity of fragments (which reflects the conservation of the site and probably the conservation of some structural aspects of it) as an important criterion for the characterization. This allows to consider the characterization of positions according to their context. Protomata-L introduces also the possibility to produce discriminative characterization of a set of sequences with respect to another one. With regard to Grammatical Inference, the confrontation of the classical state-merging techniques with a real application has lead to a new approach based on merging similar fragments. The sole application specific parts are the first and the last step of our approach (the selection of the SFPs step and the physico-chemical properties identification step) and could be replaced by similar modules for other applications. All the remaining of the approach is generic and we expect it to be an inspiration source for new theoretical or algorithmic developments. Among the originalities with respect to the classical approaches, we would like to point out the consideration of the similarity between the symbols of the alphabet, the choice of the non-deterministic representation of automata, the use of fragment-based heuristic to infer this kind of models, the identification of informative positions and the discriminative setting with respect to counter-examples (or unlabeled set of sequences) which replaces the classical compatibility setting and allows to handle some noisy counter-examples.

References

1. Hulo, N., Sigrist, C.J.A., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A.: Recent improvements to the PROSITE database. Nucl. Acids Res. **32** (2004) D134–137
2. Rigoutsos, I., Floratos, A., Parida, L., Y.Gao, Platt, D.: The emergence of pattern discovery techniques in computational biology. Metabolic Engineering **2** (2000) 159–177
3. Brejova, B., DiMarco, C., Vinar, T., Hidalgo, S., Holguin, G., Patten, C.: Finding Patterns in Biological Sequences. Unpublished project report for CS798G (2000)
4. Jonassen, I., Collins, J., Higgins, D.: Finding flexible patterns in unaligned protein sequences. Protein Science **4** (1995) 1587–1595
5. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics **14** (1998) 55–67
6. Califano, A.: Splash: structural pattern localization analysis by sequential histograms. Bioinformatics **16** (2000) 341–357
7. Eddy, S.: Hmmer user's guide: biological sequence analysis using prole hidden markov models. http://hmmer.wustl.edu/ (1998)
8. Karplus, K.: Hidden markov models for detecting remote protein homologies. Bioinformatics **14** (1998) 846–865
9. Oncina, J., Garcia, P.: Inferring regular languages in polynomial update time. Pattern Recognition and Image Analysis (1992) 49 – 61
10. Lang, K.J.: Random dfa's can be approximately learned from sparse uniform examples. 5th ACM workshop on Computation Learning Theorie (1992) 45 – 52
11. Sakakibara, Brown, Hughey, Mian, Sjolander, Underwood, Haussler: Recent methods for RNA modeling using stochastic context-free grammars. In: CPM: 5th Symposium on Combinatorial Pattern Matching. (1994)
12. Nevill-Manning, C., Witten, I.: Identifying hierarchical structure in sequences: A linear-time algorithm. Journal of Artificial Intelligence Research **7** (1997) 67–82
13. Yokomori, T.: Learning non-deterministic finite automata from queries and counterexamples. Machine Intelligence **13** (1994) 169–189
14. Coste, F., Kerbellec, G.: A similar fragments merging approach to learn automata on proteins. Technical report, IRISA, PI-1735 (2005)
15. Morgenstern, B.: DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15** (1999) 211–218
16. Lerman, I., Azé, J.: Indice probabiliste discriminant de vraisemblance du lien pour des données volumineuses. RNTI-E-1, numéro spécial Mesures de Qualité pour la Fouille des Données, H. Briand, M. Sebag, R. Gras, F. Guillet, CEPADUES (2004) 69–94
17. Lang, K.J., Pearlmutter, B.A., Price, R.A.: Results of the abbadingo one DFA learning competition and a new evidence-driven state merging algorithm. Lecture Notes in Computer Science **1433** (1998) 1–12
18. Coste, F., Fredouille, D.: What is the search space for the inference of nondeterministic, unambiguous and deterministic automata ? Technical report, IRISA - INRIA, RR-4907 (2003)
19. Taylor, W.R.: The classification of amino acid conservation. Journal of theoretical Biology **119** (1986) 205–218
20. Karkouri, K.E., Gueune, H., Delamarche, C.: Mipdb: a relational database dedicated to mip family proteins. Biol Cell **97** (2005) 535–543
21. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89** (1992) 10915–10919