Evolving Sequence Patterns for Prediction of Sub-cellular Locations of Eukaryotic Proteins

Gregory Paperin

Monash University, Faculty of Information Technology
Monash University Clayton Campus, Building 63

Wellington Road, Clayton, 3800 Victoria, Australia

gpaperin@infotech.monash.edu.au

ABSTRACT

A genetic algorithm (GA) is utilised to discover known and novel PROSITE-like sequence templates that can be used to classify the sub-cellular location of eukaryotic proteins. While traditional machine learning techniques present a black-box approach to this problem, the current method explicitly represents the discovered localisation motifs. A combined multi-class location classifier is presented and compared to other techniques based on genetic programming. Without consideration of additional structural information the presented method outperforms the alternative techniques.

Categories and Subject Descriptors

J.3. [Computer applications]: Life and Medical Sciences – *biology and genetics*. I.2.6 Learning

General Terms:

Algorithms, Experimentation

Keywords

Genetic Algorithm, Protein Localisation, Classifier Learning.

1. INTRODUCTION

The common approach to determining the location of a protein within a cell is to use sequence motifs – sequence templates typical of proteins of a specific sub-cellular location. However, this is not always satisfactory as newly discovered proteins may not contain any of the known motifs, but some others, which have not yet found their way into the databases.

A number of methods exist to predict the location of eukaryotic proteins within the cell. Hidden Markov Models [12], Neural Networks [1, 10] and Support Vector Machines [5] are widely used. Although these approaches are based on well understood mathematical models they share one drawback: they are based on a black-box approach and it is difficult to make sense of the internal rules such classifiers use. Another approach is to use a genetic algorithm (GA) to automatically evolve sequence patterns for proteins found in different sub-cellular locations. GAs and related techniques were previously used to automatically infer location motifs for eukaryotic proteins [3, 6].

The preliminary work presented here explores an alternative approach. Classifiers for the following four types of proteins are evolved. Proteins located in: (A) the cytosol, (B) inside of the nucleus, (C) inside of mitochondria and (D) extracellular proteins.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

Tested on a set of data not used during the learning process, the sensitivity of the classifiers was determined to be approximately 40% and the specificity approached 95%.

2. METHOD

A classifier-pattern is an ordered list of items that can match specific amino acids. Four different types of items are used here:

- Amino Acid: This item describes a specific amino acid.
- Gap: Matches against any amino acid.
- Property: Describes one of the amino acid properties small, hydrophobic, polar, positive, negative, tiny, aliphatic and aromatic.
- Group: Groups an arbitrary list of amino acids and matches against any amino acid in the group.

A pattern matches a protein when the protein sequence contains a subsequence of amino acids that is exactly matched by a subsequence of the pattern. The aim is to evolve patterns that match proteins found in specific sub-cellular locations. An elitist GA with a promotion rate of 20% and a 3-tournament selection is used to evolve a population of 50 patterns over 1000 generations. The initial population of patterns consists of random patterns, whose lengths are uniformly distributed between 1 and 70 items. The fitness of a classifier-pattern for a particular sub-cellular location is determined using the MCC coefficient [8]:

$$fitness = \frac{tp \times tn - fp \times fn}{\sqrt{(tn + fn) \times (tn + fp) \times (tp + fn) \times (tp + fp)}}$$

Here, tp denotes the number of true positives, tn – the number of true negatives, fp – false positives, and fn – false negatives.

The confidence c of a classifier (estimated probability for the classification to be correct) is determined using a separate data set. For positive classifications: $c = sensitivity = tp / (tp + fn)$. For negative classifications: $c = specificity = tn / (tn + fp)$.

The GA uses the following genetic operators:

Crossover: A standard GA-crossover with probability 0.85.

Elongation: At any position of a pattern a new random item is inserted with the probability 0.0005.

Mutation: Each item in a sequence is mutated with the probability 0.1. The items are mutated differently depending on their type. The semantics of the mutation operator are inspired by the various processes that can lead to mutations in eukaryotic proteins. Each type of a pattern-item can be removed from or duplicated in the sequence. *Amino acid items* can in addition be replaced by another amino acid item probabilistically chosen on the basis of the BLOSUM62 [4] substitution matrix, or swapped

for an appropriate property or group-item. *Gap items* can be swapped for an amino acid or a group-item. *Property pattern-items* can be swapped for a different property or for a group-item containing predominantly amino acids that exhibit the property described by this item. *Group items* can be modified by adding or removing group members. Groups can also be replaced by an appropriate property item. Short groups can be replaced by an amino acid or a gap and long groups can be split in two.

The described approach is implemented using an open-source GA experimentation engine JAGA [9]. This engine was chosen as all of the required standard GA operators were readily available as part of the engine and due to the convenience with which the specific operators required here could be implemented using the JAGA API.

3. EXPIREMENTAL DATA

Classifiers are evolved for proteins of four types defined according to their sub-cellular location:

A) Cytosol	C) Mitochondria
B) Nucleus	D) Extracellular proteins

Each classifier is evolved independently of the other classifiers to discriminate between proteins of a certain type $T \in \{A, B, C, D\}$ against the proteins of all other types.

Throughout the experiments a set of 1331 non-homologous proteins is used. A separate set of 145 proteins is used for validation. The evolution process is started with 20% of the training set selected at random. As the fitness of the best individual in the population increased, the rest of the training set is added in 10%-steps.

4. RESULTS

The evolved classifiers are tested on the validation data. The results are summarised in table 1.

Using these classifiers, a combined classifier is constructed based on a generalisation of the Dempster-Shafer theory [11].

Here, the belief and the plausibility that a tested protein belongs to a type T are based on the results of applying each of the four classifiers to that protein. The confidence of each classification result (i.e. sensitivity or specificity values respectively) is used as a measure of evidence that the protein belongs or does not belong to type T. The evidence is collected from all classifiers and then normalised.

Table 1. The sensitivity, the specificity and the MMC values of each evolved classifier evaluated on the validation dataset.

Classifier	Type	Sensitivity	Specificity	MCC
1	A vs. (B ∨ C ∨ D)	26.47%	98.20%	0.4206
2	B vs. (A ∨ C ∨ D)	32.88%	86.11%	0.2546
3	C vs. (A ∨ B ∨ D)	30.43%	90.16%	0.2314
4	D vs. (A ∨ B ∨ C)	60.00%	98.46%	0.6650

5. DISCUSSION & FUTURE WORK

The above results can be compared to [3], where a different GA-based approach is used to classify nucleic proteins.

The method discussed here outperforms the results shown in [3] for all cases where no additional information about secondary structure is used. However, when additional information about protein structure is used to aid the evolutionary learning [3], the resulting classifier outperforms the current approach.

In order to help the evolution to start of, patterns for known localisation motifs can be added to the initial population. In an initial set of experiments, some nuclear localisation signal motifs [2, 7] are added to the initial population used for the evolution of the classifier "B vs. (A ∨ C ∨ D)". This increases the MOC coefficient value for the training set from 44% to 51%. Further studies are required to investigate the impact of this modification on the classification of the validation data for all four classifiers.

Overall, the results show that the GA-approach to sub-cellular localisation of eukaryotic proteins has the potential to compete with more traditional methods. When extended to support all PROSITE-style regular expressions, the current strategy may discover new localisation motifs in known and unknown proteins.

6. REFERENCES

[1] Cai, Y.D. and Chou, K.C. Using Neural Networks for Prediction of Subcellular Location of Prokaryotic and Eukaryotic Proteins. *Molecular Cell Biology Research Communications*, *4* (3). 172-173.

[2] Christophe, D., Christophe-Hobertus, C. and Pichon, B. Nuclear targeting of proteins how many different signals? *Cellular Signalling*, *12* (5). 337-341.

[3] Heddad, A., Brameier, M. and MacCallum, R.M. Evolving Regular Expression-Based Sequence Classifiers for Protein Nuclear Localisation. *Applications Of Evolutionary Computing: EvoWorkshops 2004: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, and EvoSTOC, Coimbra, Portugal, April 5-7, 2004; Proceedings*.

[4] Henikoff, S. and Henikoff, J.G. Amino Acid Substitution Matrices from Protein Blocks. *PNAS*, *89* (22). 10915-10919.

[5] Hua, S. and Sun, Z. Support vector machine approach for protein subcellular localization prediction, Oxford Univ Press, 2001, 721-728.

[6] Koza, J.R., Bennett, F. and Andre, D. Using programmatic motifs and genetic programming to classify protein sequences as to extracellular and membrane cellular location. *Evolutionary Programming VII: Proceedings of the 7th Annual Conference on Evolutionary Programming*.

[7] Macara, I.G. Transport into and out of the nucleus. *Microbiol. Mol. Biol. Rev*, *65* (4). 570–594.

[8] Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Biophysica Acta*, *405* (2). 442-451.

[9] JAGA - Java API for Genetic Algorithms. `http://www.jaga.org`

[10] Reinhardt, A. and Hubbard, T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, *26* (9). 2230-2236.

[11] Shafer, G. A Mathematical Theory of Evidence. *Princeton, NJ*.

[12] Yuan, Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Letters*, *451* (1). 23-26.