

Evolving discriminative motifs for recognizing proteins imported to the peroxisome via the PTS2 pathway

Mikael Bodén and John Hawkins
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia.

Abstract—Peroxisomes are small subcellular compartments that utilize proteins manufactured in the cytoplasm. Proteins use one of two peroxisomal import pathways. This paper presents a simple evolutionary search for a motif that describes the signal used by one of the two pathways: PTS2. The evolved motif has a discriminative accuracy exceeding previously manually curated motifs and can be used to screen genomic data for putative peroxisomal proteins.

I. INTRODUCTION

Proteins in the subcellular peroxisome are nuclear encoded, translated on free ribosomes and then rely on two known import pathways. Each pathway is associated with a distinct peroxisomal targeting signal (PTS), labeled PTS1 and PTS2. Peroxisomes play an essential role in lipid metabolism for almost all eukaryotes. Deficient targeting to the human peroxisome is associated with serious diseases (e.g. Zellweger syndrome) and has prompted efforts to understand the underlying import mechanisms.

A few PTS1 predictors have been developed based on data-driven analysis, including PeroxiP [5], PTS1 [11], and PTS1Prowler [8]. However, the lack of PTS2 data has so far inhibited data-driven analysis of the second pathway into the peroxisome. This paper applies machine learning methodology to shed light on the composition of PTS2. More specifically, by utilising evolutionary search operators, we identify a discrete motif description (a pattern of amino acids) that can be used to discriminate proteins that will be imported into the peroxisome using the PTS2 from those that will not.

With 20 natural amino acids, and with a target motif of length 9, and then allowing one to 20 amino acids in each position, an exhaustive search for an optimal motif is computationally infeasible. Instead, inspired by natural evolution of biological sequences, we create a population of random candidate motif definitions based on possible amino acids in each of the nine positions. Armed with a small set of experimentally confirmed non-PTS1 targeted peroxisomal proteins, each of the candidate definitions is evaluated in terms of its discriminative ability. In contrast to conventional motif-finding algorithms we make use of a large set of known negative sequence data. Highly ranked candidates are used for creating a new population of candidate definitions using recombination and mutation.

Using evolutionary search operators we are able to identify a novel, and completely transparent motif definition for

peroxisomal targeting signal 2. The new motif is used to predict PTS2s in sequence data with a better classification accuracy than any of the previous, alternative definitions. The motif's generalization abilities are rigorously evaluated using cross-validation. We show that understanding a motif's limitations may also assist in large-scale analysis.

II. BACKGROUND

Recent studies have illustrated the two peroxisomal import pathways' usage of a set of peroxins (chaperone proteins involved in peroxisomal biogenesis). Proteins with a peroxisomal targeting signal 1 bind the receptor peroxin Pex5 and then docks on the single-lined membrane populated by a complex of other peroxins [1]. The PTS1 is carried by a majority of peroxisomal proteins and has been relatively well-studied. PTS1 is typically described as a C-terminal tripeptide (often the amino acids 'SKL') with subtle dependencies ranging the preceding nine residues [10].

Proteins with a peroxisomal targeting signal 2 first interacts with the receptor peroxin Pex7. With some species-variation, Pex7 usually associates with Pex5 before they dock on the peroxisomal membrane, unloading the cargo protein into the peroxisomal matrix [1]. The PTS2 signal (enabling the interaction with Pex7) is generally described as a nine-amino acid pattern.

In the original experiments leading to the discovery of the PTS2 it was noticed that in many peroxisomal proteins without a PTS1 motif there was an N-terminal region with many of the features of a mitochondrial targeting peptide: a net positive charge, clustering of hydroxylated residues, and the absence of a stretch of hydrophobic residues [12]. Although usually N-terminal, and often cleaved in mammalian cells, it can be found anywhere in the protein. A specific nine-amino acid consensus sequence was identified RLXXXXX[HQ]L [4], where all letters symbolize a specific amino acid (except X which stands for any amino acid) and brackets indicate optional amino acids (in a regular expression fashion). Mutational analysis has shown certain residues (in particular the first R and eighth H) to be critical [7]. This specific motif was softened to allow certain property preserving substitutions. For some years the loosest consensus sequence for PTS2 proteins has been [RK][LVI]XXXXX[HQ][LA] [14], [13].

Needless to say, motifs are over-simplified depictions of the required properties of a genuine PTS2. However, the transparency of their representation simplifies manual

screening of novel proteins as well as expert scrutiny on the grounds of additional and more specific evidence. For instance, the importance of structural properties of PTS2-imported peroxisomal proteins is yet to be fully understood [6], [3] and could be considered on top of the motif in a specific biological context.

In recent years more refined versions of the PTS2 motif have appeared. Petriv *et al.* give the most recent and specific definition of the PTS2 which they base on an analysis of known PTS2 proteins. They identified motifs that narrow down some of the central five residues. They present a highly specific version R[LVIQ]XX[LVIH][LSGA]X[HQ][LA], and a more general version [RK][LVIQ]XX[LVIHQ][LSGAK]X[HQ][LAF]. However, they note that the internal structure only emerges in a composition analysis when they focus on the N-terminal motifs, indicating that the structure of the motif may indeed have some dependence on its location within the sequence [13].

III. DATA

As SWISSPROT lacks a specific annotation for PTS2, we created a set of potential PTS2 targeted proteins by filtering SWISSPROT release 48.8 for all proteins annotated with peroxisomal localization (in the CC field), but lacking a “microbody targeting signal” (used for PTS1) or membrane association (peroxisomal membrane proteins seem to be handled by different mechanisms altogether [1]). We excluded proteins whose location was annotated as “probable”, “by similarity” or “potential.” This procedure ensured we had the current list of experimentally determined peroxisomal matrix proteins with no known import mechanism. In the absence of evidence to the contrary, we thus assume the protein to be imported by means of PTS2. The initial filtering resulted in 109 proteins.

We define a very non-specific PTS2 motif imposing necessary, but not sufficient, conditions on a PTS2 (see Table I). This “baseline” motif essentially defines a superset of all sets defined by previous PTS2 motifs. Using the non-specific baseline motif, we identified 12 of the 109 proteins as not utilising a PTS2 after all and these were consequently removed.¹ To remove homologous proteins from further consideration (and reduce bias in our predictions) we performed redundancy reduction on the 97 protein sequences,² retaining a set of 73 non-homologous samples, and used as our final set of positives.

In addition to the baseline motif, we use a hierarchy of PTS2 motifs from the literature. Each motif is strictly more specific than the other (see Table I). Due to the permissive nature of the baseline motif, several sequences have more than one match of it. We rely on this hierarchy to find the instance most likely to be a PTS2, i.e. the nine-amino acid segment that fulfils the requirement of the strictest of the

¹Of the 12 proteins, eight could be removed outright from closer inspection of the SWISSPROT entries (either incompletely specified or erroneous).

²We used BlastClust with default settings to perform redundancy reduction.

four candidate motifs. A sequence logo of the region around the PTS2 motif is shown in Figure 1. The logo illustrates the general composition of the nine amino acids (identified as per above).

To generate negatives we extracted all proteins with an experimentally confirmed and unambiguous localization either to the Nucleus, Chloroplast, Mitochondria or the Cytoplasm from SWISSPROT.³ Redundancy reduction was performed such that no two proteins have more than 20% sequence identity. Similar to the positive set, we included only proteins which matched the baseline motif. In fact, by mere chance it appears, the permissive baseline motif matches close to every second protein in SWISSPROT. We also included known PTS1 proteins that matches the baseline motif. This was done to include proteins that can be functionally similar to the PTS2 proteins yet use a different import mechanism. The total number of proteins in the final negative set is 2799.

The corresponding sequence logo graph for negative data is provided in Figure 2. Again, when more than one match is found, the nine-amino acid segment matching the strictest of the four motifs is used. The logo graphs of positive and negative data illustrate clearly that the task of differentiating them is not trivial, and that there is subtle conservation in positions 3-7 in the positive data set not seen for negatives.

IV. METHOD

We investigate the use of motifs as discriminative functions and aim to develop a novel motif that improves on reported motifs in terms of classification accuracy as viewed on novel samples.

It has been suggested that the current definition of a PTS2 motif is too broad [13]. Indeed, using the standard motif [RK][LVI]XXXXX[HQ][LA] [14] to identify positives (and to regard all others as negatives) there will be a large number of false positives, i.e. hits which are not actually recognized by the PTS2 import machinery.

The four motifs in our hierarchy (Table I) were tested for their discriminative ability. We scanned all sequences (73 positives and 2799 negatives) using each motif. For each motif, the number of sequences with at least one matching nine-amino acid sub-sequence, was recorded, as was the number of sequences not matched at all. The counts are provided in Table II.

Accuracy can not be justly quantified by a simple percentage since the two classes are highly unbalanced. Matthews’ correlation coefficient (MCC) [9] takes into account the numbers of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) of a class (Equation 1).

$$\frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \quad (1)$$

³We exclude proteins that are associated with the secretory pathway as they are localized co-translationally and would thus not be subject to peroxisomal import mechanisms even if they had valid PTSs.

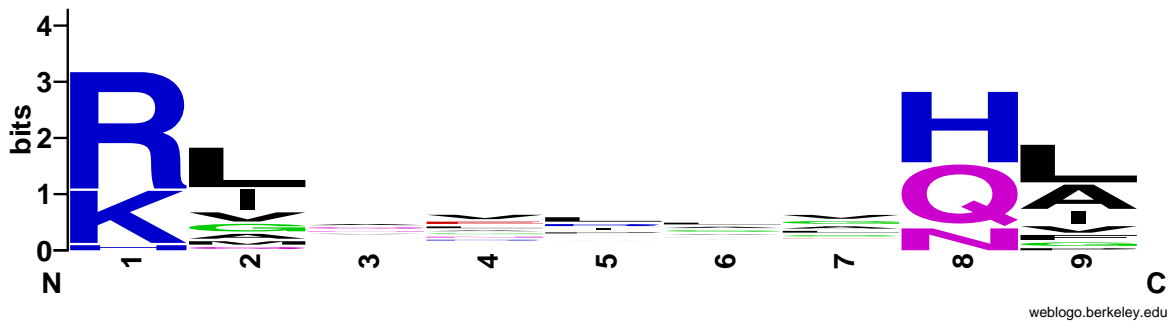


Fig. 1. A sequence logo of the aligned region containing the strongest PTS2 motif for all positive samples. Each letter stands for an amino acid prevalent in the position (x axis), the height (y axis) indicates the information theoretic content carried by the amino acid (bits).

Hierarchy of PTS2 motifs

Motif	Source
R[LVIQ]XX[LVIH][LSGA]X[HQ][LA]	[13] (specific version)
[RK][LVIQ]XX[LVIHQ][LSGAK]X[HQ][LAF]	[13] (general version)
[RK][LVIQ]XXXXXX[HQ][LAF]	[14]
[RKH][GALVIQPM]XXXXXX[HQN][LAGVIFPM]	Baseline

TABLE I

PTS2 MOTIFS IN THE LITERATURE (RANGING THE MOST TO THE LEAST SPECIFIC).



Fig. 2. The sequence logo of the aligned region containing the strongest PTS2 motif for all negative samples.

Moreover, to enable a full comparison with alternative methods we show the *sensitivity* (Equation 2),

$$\frac{tp}{tp + fn} \quad (2)$$

and the *specificity* of each class (Equation 3) for each discriminative motif.

$$\frac{tn}{tn + fp} \quad (3)$$

In order to optimize the prediction accuracy of the motif, we can not simply find an optimal fit. The best fit over all known data would not necessarily generalize well, e.g. a look-up table with all known PTS2 sequences would not extrapolate, it would simply require positive novel samples to be exactly the same as at least one already known sequence. To properly test for generalization, we generate motifs from a training set which is then evaluated on a separate test set.

We use cross-validation, whereby we divide the existing data in k sub-sets. The subsets are shuffled between each of k repeats of the simulations so that every single sequence ends up as a test sequence in exactly one run.

Can we come up with a better motif by using known non-PTS2 sequences? That is, can we find a better classification boundary between positives and negatives by exploring the negatives that are near in sequence-space? We attempt to answer this question by evolving a discriminative motif under selective pressure defined over both positive and negative data. For each sequence we save only the nine consecutive amino acids that match the most specific motif of the four in Table I. Thus, we base our selective criteria on basis of these 73+2799 9-amino acid sequence segments only. We believe that searching by evolutionary operators may assist in identifying candidate motifs that are less vulnerable to accidental mutations, i.e. motifs that are less likely to classify

Classification results for PTS2 motifs

Motif	TP	FP	TN	FN	Sensitivity	Specificity	MCC
[13] (specific)	17	37	2762	56	0.23	0.99	0.25
[13] (general)	20	112	2687	53	0.27	0.96	0.18
[14]	38	709	2090	35	0.52	0.75	0.10
Baseline motif	73	2799	0	0	1.00	0.00	NaN

TABLE II

DISCRIMINATIVE ABILITY OF PREVIOUSLY REPORTED MOTIFS AND OUR BASELINE MOTIF (WHICH IS USED AS A PRE-FILTER FOR ALL SCREENED SEQUENCES). PERFECT SENSITIVITY OR SPECIFICITY IS 1.0, THE WORST VALUE IS 0.0.

positives as negatives even if they are perturbed slightly.

We define a candidate PTS2 motif in the following way. There are nine positions. In each position a subset, excluding the empty set, of all allowed amino acids is identified. We use the baseline motif to decide which amino acids are allowed—all others would be futile as matching the baseline motif is a prerequisite for the data. For example, in position one, R, K and H are allowed by the baseline motif. A candidate PTS2 motif can thus in position one either have R, K, or H by themselves, RK, RH, KH, or RKH.

Initially, we randomly generate n candidate motifs (fulfilling the requirement above). Each candidate motif is evaluated using an objective (or fitness) function. In our first set of simulations the objective function is simply the MCC, over the training set. All n candidate motifs are ranked. Depending on the rank the motif is selected for recombination – the higher rank, the higher probability (as defined by the absolute of a Gaussian distribution with a mean of 0 and a variance of $n/2$). In total, $n/3$ number of motifs are selected using the rank for recombination. The offspring of recombining two selected parent motifs (from the above $n/3$ -set) is generated by first shuffling the parents’ motifs at each point between the 9 positions with a probability of $P(Recomb)$. This probability is small, thus parents usually exchange their parts only between one or two positions. Thereafter the offspring definition is subject to mutation: each part of the motif is perturbed with a probability $P(Mut)$. The probability of mutation is also small, but when it happens it consists of adding or removing an amino acid with equal probability. A check is done so that at least one amino acid is allowed in each position. Of the n motifs in the population, we keep $n/6$ unchanged (elitism) and replace the remaining $5n/6$ with the offsprings (generated as per above).

A range of objective functions were tested. As the final accuracy is measured in terms of MCC on a test set it is perhaps not surprising to see that using MCC as an objective function over the training data works well.

To compensate for the unbalanced training data, we trialed a weighted MCC objective function. The number of true positives and false negatives (i.e. all known positive samples) was multiplied by a weight which was varied between simulations.

We also designed an objective function to penalize complicated motif descriptions. More specifically, we identified

a range of graded scales describing the physicochemical properties of amino acids (hydrophobicity, charge and type). Each amino acid identified one value in each of these scales. For each position in a candidate motif we calculated an entropy for each scale, depending on the accepted amino acids. For each position, the smallest entropy value over the three scales was chosen. A small entropy would thus indicate that the set of allowed amino acids had similar properties in regard to at least one scale. A large entropy would indicate that there was great variation among the values for all three scales. The objective function is the sum of the MCC and the negative sum of position-specific entropy values. We choose to control the influence of the summed entropy by means of a weight.

V. RESULTS

For all simulations we decided to use a population of $n = 30$ motifs, evolved for 300 generations. Some preliminary trials indicated that simulations usually converged at some point before 300 generations and that more than 30 motifs rarely made a difference. We also tested some variations in setting recombination and mutation probabilities and decided to report results for $P(Recomb) = 0.25$ and $P(Mut) = 0.10$ as they seemed to ensure some initial exploration and then stable convergence. From the large number of simulations we ran we saw no results that were significantly outperforming those reported here. On a cautionary note, we freely acknowledge that there may be more optimal parameter settings. However, our intention here is primarily to illustrate the composition, use and accuracy of an evolved PTS2 motif.

A. MCC as objective function

Using 10-fold cross-validation, we evolved motifs using the MCC objective function. We then repeated the simulation 10 times with different initial populations and with different divisions of the data sets. Consequently, 100 motifs were created.

The average MCC over the test data is 0.299 with a standard deviation of 0.041 over the 10 repeats. Note that in order to calculate the MCC for a single run one has to use all 10 motifs—each tested only on those samples that were excluded while training. Every single motif had an extremely high specificity indicating high integrity—almost no non-PTS2 sequences were classified incorrectly as positives. However, this seems to happen at the expense of the

sensitivity which was usually very low. All data is presented in Table III.

B. A weighted MCC as objective function

In our weighted MCC simulations we used four different weight values to compensate for the paucity of positive samples. We observed no or very variable improvement in sensitivity, a slight decrease in specificity, and a decrease in MCC as measured over the non-weighted test data (see Table III).

C. Penalizing physicochemical disorder

By penalizing sets of amino acids that had different physicochemical properties with the negative entropy objective function, we noticed that accuracy dropped dramatically beyond a weight of 0.1.

D. A final motif

The accuracy measured over the training data is generally much better than over the test data. In our tests we note that the MCC over the training data is usually between 0.50 and 0.60. However, this value does not represent the true ability of the motif when used on novel samples. However, the performance that we observe over the test data for a particular configuration is indicative of the performance that we would get from a motif evolved using that same configuration gleaning the whole data set. We choose to use the whole data set to evolve the final motif, and use the accuracy we got over a smaller training data set as an estimate of its accuracy on genuinely novel sequences.

The evolved motifs are shown in Figure 3.

E. Screening *Arabidopsis Thaliana* for PTS2

We screened the full genome of *Arabidopsis Thaliana* from NCBI [15] using the motif evolved using the simple MCC (the first motif in Figure 3). After removing all sequences with predicted signal peptides localized to the secretory pathway (using the Protein Prowler subcellular localization predictor [2]), 76 sequences were identified as using a PTS2. Since the sensitivity is believed to be about 0.2 we expect that the count is an underestimate of the true number of PTS2-imported proteins. However, due to the near perfect specificity most of the 76 proteins are likely to be true PTS2 proteins.

F. Screening the peroxisome-less *Entamoeba* for PTS2

The eukaryotic *Entamoeba* is believed to specifically lack a peroxisome and its proteins do therefore not require peroxisomal targeting mechanisms. Any instances of PTS2 signals (or PTS1 for that matter) are not under selective pressure. We expect that there would be no PTS2 signals amongst *Entamoeba* proteins except by mere chance. Hence, the *Entamoeba* genome serves as a strong test of the integrity of the evolved motif.

We extracted the protein set from the *Entamoeba* genome (again from NCBI), consisting of 20369 proteins, and scanned it for matches to the same motif which had 76

hits in *Arabidopsis Thaliana*. We found a mere 12 false positives, after filtering predicted secreted proteins, matching the expected specificity at 0.9994.

VI. CONCLUSION

We use a simple motif description to discriminate between PTS2-imported proteins and others. This representation has obvious limitations, e.g. it can not indicate graded membership or dependencies between individual positions. However, it is transparent and can thus be used in an informed and considered manner by the experimenter.

We evolve a PTS2 motif with a classification accuracy that exceeds previously proposed PTS2 motifs. We also estimate its accuracy in such a way that we can predict how many mistakes it will make. We screen the genome of *Arabidopsis Thaliana* and identify 76 putative PTS2 imported proteins. By screening the peroxisome-less *Entamoeba* genome, we confirm that the estimated specificity is very high and we thus expect very few false positives when using the motif. The estimated sensitivity is however very low and we believe that a predicted set represents only about 20-30% of the complete set of PTS2-imported proteins.

Our study not only identifies a novel PTS2 motif but also demonstrates the natural utility of evolutionary search operators in creating biological sequence motifs—which represent entities that themselves are subject to selective pressure.

REFERENCES

- [1] A. Baker and I. A. Sparkes. Peroxisome protein import: some answers, more questions. *Current Opinion in Plant Biology*, 8(6):640–647, 2005.
- [2] M. Bodén and J. Hawkins. Prediction of subcellular localisation using sequence-biased recurrent networks. *Bioinformatics*, 21:2279–2286, 2005.
- [3] C. B. Brocard, C. Jedeszko, H. C. Song, S. R. Terlecky, and Paul A. Walton. Protein structure and import into the peroxisomal matrix. *Traffic*, 4(2):74–82, 2003.
- [4] M. J. de Hoop and G. Ab. Import of proteins into peroxisomes and other microbodies. *The Biochemical Journal*, 286(3):657–669, 1992.
- [5] O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal. In silico prediction of the peroxisomal proteome in fungi, plants and animals. *Journal of Molecular Biology*, 330(2):443–456, 2003.
- [6] C. R. Flynn, R. T. Mullen, and R. N. Trelease. Mutational analyses of a type 2 peroxisomal targeting signal that is capable of directing oligomeric protein import into tobacco by-2 glyoxysomes. *The Plant Journal*, 16(6):709–720, 1998.
- [7] C. Gietl. Protein targeting and import into plant peroxisomes. *Physiologia Plantarum*, 97(3):599–608, 1996.
- [8] J. Hawkins and M. Bodén. Predicting peroxisomal proteins. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 469–474, Piscataway, November 2005. IEEE.
- [9] B. W. Matthews. Comparison of predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, 405:442–451, 1975.
- [10] G. Neuberger, S. Maurer-Stroh, B. Eisenhaber, A. Hartig, and F. Eisenhaber. Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *Journal of Molecular Biology*, 328(3):567–579, 2003.
- [11] G. Neuberger, S. Maurer-Stroh, B. Eisenhaber, A. Hartig, and F. Eisenhaber. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of Molecular Biology*, 328(3):581–592, 2003.

Classification results for evolved PTS2 motifs

Simulation	Sensitivity	Specificity	MCC
Standard MCC	0.190 (0.028)	0.995 (0.001)	0.299 (0.041)
Weighted 2x	0.192 (0.021)	0.994 (0.001)	0.291 (0.032)
Weighted 5x	0.177 (0.041)	0.995 (0.001)	0.281 (0.045)
Weighted 10x	0.185 (0.022)	0.991 (0.011)	0.265 (0.049)
Weighted 100x	0.296 (0.079)	0.896 (0.066)	0.117 (0.056)
Neg entropy 0.05x	0.196 (0.019)	0.995 (0.001)	0.298 (0.028)
Neg entropy 0.1x	0.194 (0.028)	0.995 (0.001)	0.302 (0.030)
Neg entropy 0.2x	0.341 (0.070)	0.863 (0.051)	0.098 (0.031)

TABLE III

DISCRIMINATIVE ABILITY OF EVOLVED MOTIFS. AVERAGES ARE REPORTED, STANDARD DEVIATIONS OVER 10 REPEATS IN BRACKETS.

[R][LIQM][^CHLYV] [REHILKMFVS][^AEGHKPSW][^RDEGFPW][^I] [H][LA]
 [R][LIQM][^DCHILYV][RDEHLMSWV] [^QEGK] [^GPW] [^IPW][H][LAM]

Fig. 3. The motif evolved using the plain MCC as an objective function and the motif evolved using the MCC and physicochemical disorder objective function. Caret signifies negation.

- [12] T. Osumi, T. Tsukamoto, S. Hata, S. Yokota, S. Miura, Y. Fujiki, M. Hijikata, S. Miyazawa, and T. Hashimoto. Amino-terminal pre-sequence of the precursor of peroxisomal 3-ketoacyl-coa thiolase is a cleavable signal peptide for peroxisomal targeting. *Biochemical and Biophysical Research Communications*, 181(3):947–954, 1991.
- [13] O. I. Petriv, L. Tang, V. I. Titorenko, and R. A. Rachubinski. A new definition for the consensus sequence of the peroxisome targeting signal type 2. *Journal of Molecular Biology*, 341(1):119–134, 2004.
- [14] R. A. Rachubinski and S. Subramani. How proteins penetrate peroxisomes. *Cell*, 83(4):525–528, 1995.
- [15] A. Theologis et al. Sequence and analysis of chromosome 1 of the plant *arabidopsis thaliana*. *Nature*, 408(6814):816–820, 2000.