Supporting Appendix

Software Availability. Software is available from the authors upon request.

Partition the Profile Space via Supervised Learning. A continuous profile space is partitioned into 15 sub-spaces according to the text. To optimize the boundaries of the sub-spaces, we developed a supervised learning algorithm and implemented it as a C program NW_CLUSTER. While the program can partition the profile space into any arbitrary number of sub-spaces, we initialized the number to be 15 and gave 15 starting points as the center of each sub-space, based on some pilot analysis (data available upon request) and the work of Eskin (1).

We first collected all matrices of known yeast transcription factors, as well as multiple sequence alignments of intergenic regions of four yeast species (*Saccharomyces cerevisiae*, *Saccharomyces mikate*, *Saccharomyces kudriazevii* and *Saccharomyces bayanus*) (2). Each position of these alignments is converted into a frequency vector (f_A, f_C, f_G, f_T). These frequency vectors, or profiles, are an approximation of the entire continuous profile space of length one. Fifteen profiles are chosen to be the initial centers of the 15 subspaces. They are (1.00, 0.00, 0.00, 0.00), (0.00, 1.00, 0.00, 0.00), (0.00, 0.00, 1.00, 0.00), (0.00, 0.00, 0.00, 1.00), (0.70, 0.10, 0.10, 0.10), (0.10, 0.70, 0.10, 0.10), (0.10, 0.10, 0.70, 0.10), (0.10, 0.10, 0.10, 0.70), (0.40, 0.40, 0.10, 0.10), (0.40, 0.10, 0.40, 0.10), (0.40, 0.10, 0.10, 0.40), (0.10, 0.40, 0.40, 0.10), (0.10, 0.40, 0.10, 0.40), (0.10, 0.10, 0.40, 0.40), and (0.25, 0.25, 0.25, 0.25). A subspace is defined as a collection of profiles. The similarity between any two profiles is defined as a simplified form of ALLR statistic (3):

$$ALLR = \frac{\sum_{b=A..T} \ln(f_{bi} / p_b) + \sum_{b=A..T} \ln(f_{bj} / p_b)}{2},$$

where p_b is the background frequency of the letters. The similarity between a profile and a subspace is the sum of the pairwise similarity between this profile and every profile in such subspace, weighted by the frequencies of the profiles. The similarity between two subspaces is the sum of all pairwise similarities between each profile in the first subspace and each profile in the second subspace, weighted by the frequencies of such pairs.

The algorithm initializes by calculating the pairwise similarities between all profiles in the entire space and the 15 selected centers. Each profile is assigned to one of the centers, creating an initial 15 partitions of the profile space. Then, the similarity between each profile and each subspace is calculated. For any profile, if the similarity to its assigned subspace is smaller than the similarity to another subspace, this profile is reassigned to the closest subspace. Therefore both the original subspace and the closest subspace are redefined. All the similarities between the profiles and the redefined subspaces are recalculated accordingly. This reassignment process is iterated to maximize the similarity between any profile and its own assigned subspace, while minimizing the sum of the similarities between the profile and all the rest subspaces, until no reassignment is needed. The final assignments of the profiles define the boundaries of the 15 partitions.

Estimating λ, Target Frequencies, *H* and *K*. Average Log Likelihood Ratio (ALLR) scores are log-odds scores. The property of the scoring matrix can be revealed by solving the following two equations (4, 5):

$$S_{ij} = \frac{\log(\frac{q_{ij}}{p_i p_j})}{\lambda} \qquad \dots\dots (1)$$

$$\sum_{i,j} p_i p_j e^{\lambda S_{ij}} = 1 \qquad \dots\dots (2)$$

where *i* and *j* represent two subprofile spaces, S_{ij} is the ALLR score between two subprofile spaces, p_i and p_j are densities of the subprofile spaces. λ is solved by implementing a Newton/Raphson root finding algorithm (6). The implied target frequency can then be obtained.

λ = 0.020936

Target frequency (%):

	A	C	G	T	a	c	g	t	M	R	W	S	Y	K	n
A	15.68	0.00	0.00	0.00	4.09	0.00	0.00	0.00	0.06	0.13	0.10	0.00	0.00	0.00	0.05
C	0.00	10.34	0.00	0.00	0.00	3.32	0.00	0.00	0.07	0.00	0.00	0.05	0.14	0.00	0.03
G	0.00	0.00	10.36	0.00	0.00	0.00	3.24	0.00	0.00	0.14	0.00	0.05	0.00	0.06	0.02
T	0.00	0.00	0.00	14.73	0.00	0.00	0.00	3.84	0.00	0.00	0.09	0.00	0.13	0.06	0.05
a	4.09	0.00	0.00	0.00	2.57	0.01	0.02	0.01	0.06	0.20	0.11	0.00	0.00	0.00	0.19
c	0.00	3.32	0.00	0.00	0.01	2.52	0.00	0.02	0.09	0.00	0.00	0.06	0.27	0.00	0.13
g	0.00	0.00	3.24	0.00	0.02	0.00	2.50	0.01	0.00	0.26	0.00	0.06	0.00	0.08	0.13
t	0.00	0.00	0.00	3.84	0.01	0.02	0.01	2.56	0.00	0.00	0.11	0.00	0.21	0.06	0.19
M	0.06	0.07	0.00	0.00	0.06	0.09	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.02
R	0.13	0.00	0.14	0.00	0.20	0.00	0.26	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.05
W	0.10	0.00	0.00	0.09	0.11	0.00	0.00	0.11	0.00	0.00	0.17	0.00	0.00	0.00	0.03
S	0.00	0.05	0.05	0.00	0.00	0.06	0.06	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.02
Y	0.00	0.14	0.00	0.13	0.00	0.27	0.00	0.21	0.00	0.00	0.00	0.00	0.57	0.00	0.05
K	0.00	0.00	0.06	0.06	0.00	0.00	0.08	0.06	0.00	0.00	0.00	0.00	0.00	0.12	0.02
n	0.05	0.03	0.02	0.05	0.19	0.13	0.13	0.19	0.02	0.05	0.03	0.02	0.05	0.02	0.68

Since each deputy profile also represents a distribution of A, C, G and T, the target percentage identity at the level of nucleotides can also be implied to be ≈ 70%.

The entropy of the matrix can be calculated using the following formula:

$$H = \sum_{ij} q_{ij} \log(\frac{q_{ij}}{p_i p_j})$$

For ALLR scoring matrix, H = 3.26 bits. That means that on average there is 3.36 bits of information per aligned position.

K is computed for ungapped alignments using a random walk procedure that traverses the space of possible scores from a starting position, implemented according to source code of BLAST-1.1 (W. Gish, personal communication).

Seed Design. A seed is a substring of the query profiles. All substrings of a predefined length of the query profile as well as “neighborhood profiles” (profiles that are sufficiently similar to a seed) are cataloged and used to quickly identify local similarities between a query profile and the database profiles. Probing sequences with several distinct seeds have been shown to improve search performance. Here we adapt the idea for fast sequence similarity search to fast profile similarity search.

A seed format is a string of 0s and 1s. For example, 111111 and 111000111 are both valid seed formats. The length of the string is called “span” of the seed, and the number of 1s in the string is called “weight” of seed (7). In a seed format, 1 means that at this position we only allow an exact match; 0 means that we don’t place any requirement on this position. A seed match means that there is a substring in the database profile that is identical to a seed profile at every position that is labeled with “1.” Once a seed match is found, it is extended at both ends to identify a high scoring pair between the query profile and database profile.

By default the seed format has both a span and weight of 6. Allowing variable seed formats has several advantages: (*i*) A carefully designed seed will improve the search performance (7); (*ii*) Certain prior knowledge about transcription factor binding sites can be built into the seed. For example, zinc finger protein Gal4 binds to its DNA target site CGGN$_{11}$CCG as a dimer. The specificity is interrupted by an 11-bp nonspecific sequence region. To search for sites with a similar pattern, a seed format 111(0)$_{11}$111 can be used to incorporate such information.

Neighborhood Profile Generation. A neighborhood profile is defined as a profile with an alignment score greater than a threshold when compared to a given seed. For every seed, neighborhood profiles are generated according to its weighted positions, i.e., positions labeled with 1. The weight of a seed is important to the number of potential neighborhood profiles. For example, for a seed of weight 6, there are 11.4 million possible profiles constructed using the deputy profiles, but only a very small fraction, a hundred or so, may pass a reasonable threshold and qualify as a neighborhood word. Therefore, a branch and bound algorithm is used to generate neighborhood profiles efficiently according to the source code of BLAST-1.1 (W. Gish, personal communication).

Calculating Final *P* value of a Motif. Let $y = KMNe^{-\lambda S}$, the *P* value for finding *m* alignments with score ≥*S* is given by:

$$P = 1 - (\sum_{i=1..m} \frac{y^{m-1}}{(m-1)!})e^{-y}$$

Validation of Predicted Motifs and Gene Clusters. Another example is PHYLONET motif YCL050C.1 with consensus gGAAngTTCTAGAAg, matching the Hsf1 binding site. Three sets of genes were obtained based on PHYLONET prediction (set A, 39 genes), genome-wide computational scan (set B, 158 genes), or genome-wide location assay of Hsf1 protein under low H_2O_2 condition [set C, 68 genes, (8)]. All three sets are enriched for protein folding, and the *P* value of enrichment is 9.33E-10 for set A, 2.70E-7 for set B, and 1.87E-8 for set C (*Fig. 9A*). The expression coherence of genes in set A, B and C under DNA damage condition is 0.3463 with a *P* value of 1.23E-6, 0.064 with a *P* value of 0.694, and 0.0882 with a *P* value 0.342, respectively (Fig. 9 *B*–*D*).

1. Eskin, E. (2004) *Proc. Eighth Annu. Int. Conf. Comp. Mol. Biol.* **2004,** 115-124.
2. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) *Science* **301,** 71-76.
3. Wang, T. & Stormo, G. D. (2003) *Bioinformatics* **19,** 2369-80.
4. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460-80.
5. Altschul, S. F., Bundschuh, R., Olsen, R. & Hwa, T. (2001) *Nucleic Acids Res.* **29,** 351-361.
6. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1993) *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, U.K.).
7. Ma, B., Tromp, J. & Li, M. (2002) *Bioinformatics* **18,** 440-445.

8. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al*. (2004) *Nature* **431,** 99-104.