Volume 16 Number 5 1988 Nucleic Acids Research

# A backtranslation method based on codon usage strategy

Graziano Pesole, Marcella Attimonelli and Sabino Liuni

Dipartimento di Biochimica e Biologia Molecolare, Centro Studi Mitocondri e Metabolismo Energetico (CNR), University of Bari, Bari, Italy

Received August 17, 1987; Revised and Accepted November 15, 1987

## ABSTRACT

This study describes a method for the backtranslation of an aminoacidic sequence, an extremely useful tool for various experimental approaches. It involves two computer programs CLUSTER and BACKTR written in Fortran 77 running on a VAX/VMS computer. CLUSTER generates a reliable codon usage table through a cluster analysis, based on a chi2-like distance between the sequences. BACKTR produces backtranslated sequences according to different options when use is made of the codon usage table obtained in addition to selecting the least ambiguous potential oligonucleotide probes within an aminoacidic sequence. The method was tested by applying it to 158 yeast genes.

## INTRODUCTION

The inference of a nucleotide sequence from an aminoacid sequence is an intriguing problem from both theoretical and practical points of view, since it allows for the construction of probes for use in screening of cDNA and genomic libraries in order to isolate the corresponding gene or messenger RNA sequence. In addition, prediction of nucleotide sequences may prove useful in molecular evolution studies.

Owing to the degeneracy of the genetic code the problem of generating a reliable nucleotide sequence from an aminoacid sequence is rather complex, with approaches followed so far yielding less than satisfactory results (1,2).

This paper presents a method which generates the most likely nucleotide sequences from an aminoacid sequence, taking into account the different codon strategies adopted by the various genomes. It is in fact known that DNA coding sequences do not use the "synonymous" codons with equal frequencies, since each genome adopts its own particular strategy in codons usage (3).

The novelty of our method centres on use made of a cluster analysis based on a new formula for the calculation of the gene-distance between two sequences.

MATERIALS

The method has been realized by generating two computer programs, CLUSTER and BACKTR, written in Fortran Standard and presently used on the VAX/Cluster of the Physics Department, Bari University (Italy). CLUSTER and BACKTR require respectively input files of nucleotide and aminoacid sequences which were extracted from merged GenBank and EMBL databases as well as from NBRF database through our ACNUC software (4). Backtranslated sequences were stored in files for further analysis.

METHOD

As a first step we extracted a set of nucleotide sequences from our database, coding for proteins belonging to the same species of the polipeptide to be backtranslated. If the number of elements constituting the set is not significantly large (<1000 codons), the set may be formed by a pool of coding sequences belonging to closely related species. We carried out hierarchical clustering on this set based on similarity in codon usage, generating one or more groups of sequences homogeneous for codon strategy. The subset of which our polipeptide sequence may be a part was chosen on the basis of such biological properties as protein class (e.g. histone or ribosomal protein) or high or low expression level. For this subset we constructed a codon usage table reporting mean values of the frequencies within its family and standard deviations for each codon, implementing procedures via the CLUSTER program.

This cluster analysis required a similarity matrix containing the distances between each pair of sequences of the initial pool to assure that neighboring messengers corresponded to sequences having similar codon usage. The distance computation between the i and j sequences involved a chi2-like formula:

$$\chi^2_{ij} = \sum_{k=1}^{61} \left[ \frac{(n_i^{(k)} - E_{ij}^{(k)})^2}{E_{ij}^{(k)}} + \frac{(n_j^{(k)} - E_{ji}^{(k)})^2}{E_{ji}^{(k)}} \right] \qquad [1]$$

where $n_i^{(k)}$ is the total number of occurrences of the codon k in the sequence

i. $E_{ij}^{(k)}$ is the expected value for occurrences of codon k in the i sequence weighed by the probability $p_i^{(h)}$ that the family h, formed by c codons, occurs in i sequence in relation to j sequence:

$$p_i^{(h)} = \frac{N_i^{(h)}}{N_i^{(h)} + N_j^{(h)}} \qquad \text{with} \qquad N_i^{(h)} = \sum_{k=1}^{c} n_i^{(k)} \qquad [2]$$

$$E_{ij}^{(k)} = (n_i^{(k)} + n_j^{(k)}) \cdot p_i^{(h)} \qquad [3]$$

Since the chi2-like distance [1] does not take into account the number of independent variables in both sequences, the formula was corrected by dividing by the number of degrees of freedom $NDF_{ij}$, where:

$$NDF_{ij} = \sum_{h=1}^{20} \left[ F_{ij}^{(h)} .(DCF^{(h)})-1 \right] \qquad [4]$$

$F_{ij}^{(h)}$ being 1 or 0 depending on whether or not the h-th family is present in both sequences and $DCF^{(h)}$ is the number of degenerate codons in the h-th family. By correcting for this, the distance formula to be employed in the similarity matrix becomes:

$$D_{ij} = \chi_{ij}^{2} \, / \, NDF_{ij} \qquad [5]$$

It is thus evident that our approach for distance metrics has taken into account the following factors:

a) the degree to which total distance between two sequences depends on the number of codon occurrences in the family,

b) the aminoacidic composition of the two sequences (see [4] and [5] ).

The hierarchical complete-linkage clustering works as follows. Given the ND data points (data cluster) that in our case represent sequences in the starting pool, a new cluster is formed by combining the two nearest points. Distances were recorded and the similarity matrix updated by calculating the new distances between the new cluster and the remaining data. Updating was carried out so that the distance between two cluster corresponded to that of the two extreme points. This "updating operation" went on until all ND data points comprised a single cluster with a total of ND-1 clusters created. These ND-1 clusters were given the number ND+k, with k varying from 1 to ND-1. The result of clustering is graphically represented by a bynary tree so one or more groups of sequences which use a similar codon strategy may be distinguished. These groups constitute homogeneous subclusters for which we define "cluster density" as the ratio between the number of subcluster data points and the volume of the hypothetical sphere containing them, whose radius is calculated as the distance between the two extreme points of the subcluster $d_s^{max}$:

$$\rho_s = n_s \; / \; \frac{4}{3} \left( \frac{d_s^{max}}{2} \right)^3 \qquad [6]$$

Among these groups we selected the group (homogeneous pool) within which our polipeptide sequence may be realistically situated.

Detection of the outliers in each homogeneous pool can be performed by either direct observation of the cluster or simple chi-square statistic supplied optionally by our CLUSTER program. The chi-square statistic involves application of formula [1] for the i-th sequence when the j-th sequence is the supergene obtained linking all genes belonging to the homogeneous pool. In this case, any threshold value may be imposed.

The "center of gravity" of the homogeneous group is given by the mean values of the frequencies $\bar{f}_k$ for each degenerate codon k in the family, and provides the codon-usage table for the backtranslation of the polipeptide sequence:

$$\bar{f}_k = \sum_{i=1}^{ND} n_i^{(k)} \; / \; \sum_{i=1}^{ND} N_i^{(h)} \qquad [7]$$

In addition to mean values, the program also yields standard deviations:

$$\sigma_k = \sum_{i=1}^{ND} \left[ w_i^{(h)} \cdot (n_i^{(k)} - \bar{f}_k \cdot N_i^{(h)})^2 \right]^{\frac{1}{2}} \qquad [8]$$

$$w_i^{(h)} = N_i^{(h)} \; / \; \sum_{i=1}^{ND} N_i^{(h)} \qquad [9]$$

In addition CLUSTER may also be used to compute the distance between a sequen ce and a given pool by comparing their codon usage frequencies according to the chi-square statistics.

BACKTR performs the backtranslation of a polipeptide sequence requiring the codon-usage table generated by CLUSTER, with the following options available:

1) generation of the "most likely" sequence, choosing the codon with the highest codon usage frequency for each family;
2) generation of one or more randomly weighed sequences, constructed by assigning the codon to each aminoacid according to a Montecarlo simulation procedure based on the codon usage table;
3) generation of the ambiguous sequence with the degenerate code corresponding to IUPAC-IUB recommendations;
4) generation of completely random sequences.

BACKTR also allows for selection of the least ambiguous oligonucleotide probe, with a length of L codons. By scanning the entire sequence of NC codons with a window of L codons, the program calculates for each window (j) the product ($P_j$) of the highest $f_i^{max}$ codon usage frequencies for L codons and selects the best oligonucleotide as the one with the highest product $P_j^{max}$:

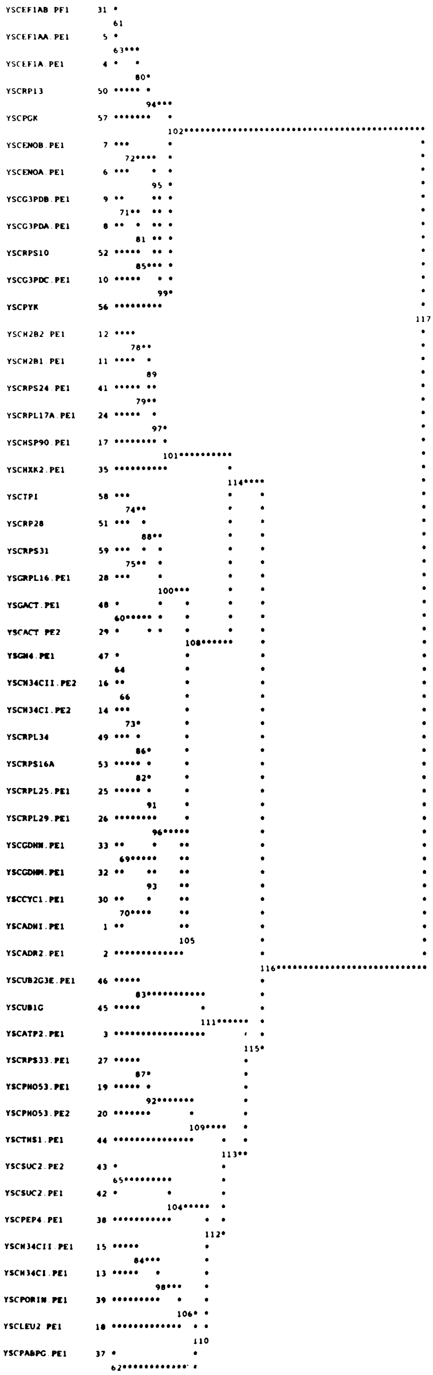
$$P_j = \prod_{i=j}^{j+L-1} f_i^{max} \qquad [10]$$

$$P_j^{max} = \mathrm{Max}\ (P_1,\ldots,P_n) \text{ where } n = NC-L+1 \qquad [11]$$

One or more randomly-weighed probes may also be obtained, as indicated in option 2.

## RESULTS

Our method was applied to 158 yeast genes extracted from our ACNUC database (release 50 GenBank and 10 EMBL). Our cluster analysis identified two groups, in agreement with Sharp, who also used yeast genes for cluster analysis (5). The first group can be further split into two subgroups made up respectively of 12 and 47 genes, including genes thought to be highly expressed, for example all the ribosomal protein genes, the dendrogram for which is shown in fig. 1. The second group includes lowly expressed genes.Tables 1 and 2 report


YSCEF1AB.PE1 31
61
YSCEF1AA.PE1 5
63
YSCEF1A.PE1 4
80
YSCRP13 50
94
YSCPGK 57
102
YSCENOB.PE1 7
72
YSCENOA.PE1 6
95
YSCG3PDB.PE1 9
71
YSCG3PDA.PE1 8
81
YSCRPS10 52
85
YSCG3PDC.PE1 10
99
YSCPYK 56
117
YSCH2B2.PE1 12
78
YSCH2B1.PE1 11
89
YSCRPS24.PE1 41
79
YSCRPL17A.PE1 24
97
YSCHSP90.PE1 17
101
YSCHXK2.PE1 35
114
YSCTPI 58
74
YSCRP28 51
88
YSCRPS31 59
75
YSGRPL16.PE1 28
100
YSGACT.PE1 48
60
YSCACT PE2 29
108
YSGH4.PE1 47
64
YSCH34CII.PE2 16
66
YSCH34CI.PE2 14
73
YSCRPL34 49
86
YSCRPS16A 53
82
YSCRPL25.PE1 25
91
YSCRPL29.PE1 26
96
YSCGDHN.PE1 33
69
YSCGDHM.PE1 32
93
YSCCYC1.PE1 30
70
YSCADH1.PE1 1
105
YSCADR2.PE1 2
116
YSCUB2G3E.PE1 46
83
YSCUB1G 45
111
YSCATP2.PE1 3
115
YSCRPS33.PE1 27
87
YSCPHO53.PE1 19
92
YSCPHO53.PE2 20
109
YSCTHS1.PE1 44
113
YSCSUC2.PE2 43
65
YSCSUC2.PE1 42
104
YSCPEP4.PE1 38
112
YSCH34CII.PE1 15
84
YSCH34CI.PE1 13
98
YSCPORIN.PE1 39
106
YSCLEU2.PE1 18
110
YSCPABPG.PE1 37
62

locus names and the short description of the first and second group. Codon usage frequencies and corresponding standard deviations for these groups are shown respectively in table 3 and 4. It is clearly evident that the codon strategy is much more biased in the group of highly expressed genes. Fig. 2 reports the cluster density map, with 5 genes observable: ATPase beta subunit, ribosomal protein S33, pre-invertase, Pho 3 and Pho 5, all located in the first group. This observation is in disagreement with Sharp's data (5) which localizes these genes in the second group. In order to verify the accuracy of our clusterization, the above-mentioned 5 genes were first translated into aminoacid sequences, then backtranslated by using the codon usage table for both highly and lowly expressed genes (table 3 and 4). Comparison between backtranslated and real sequences was made by applying the BESTFIT program of our GLORIA software (submitted to CABIOS). Results agree with our clusterization since randomly weighed sequences backtranslated according to the codon usage of table 1 gave a percentage identity markedly superior to that obtained with codon usage in table 2 (see table 6). For the sake of example, we produced the most likely probe for seven highly expressed yeast genes according to the codon usage table 1. The number of matches as compared to real nucleotide sequences is reported in table 7.

## DISCUSSION

To date relatively few backtranslation techniques have been described (1,2), the major difficulty in obtaining more realistic backtranslation being definition of a correct codon-usage table. In order to solve this problem, were taken into account the following observations: 1) DNA coding sequences do not use "synonymous" codons with equal frequencies; 2) sequences belonging to the same species adopt codon strategies that are closer than those adopted by

Fig. 1 Cluster analysis dendrogram for the 59 highly expressed genes. The horizontal length of branches are scaled according to maximum distance [5] between two groups when clustered. For each gene in the cluster the locus name is reported assigned respectively to the sequences from the GenBank (release 50 ) and EMBL (release 10 ) databases. All the clusters are progressively numbered from 60 to 117.

Tab. 1 List of the 59 highly expressed yeast genes clustered in the first group extracted from the merged GenBank (release 50) and EMBL (release 10) databases.

| locus name | description | codons |
|---|---|---|
| YSCADHI.PE1 | ALCOHOL DEHYDROGENASE I | 20 |
| YSCADR2.PE1 | ALCOHOL DEHYDROGENASE II | 349 |
| YSCATP2.PE1 | ATPASE BETA SUBUNIT | 313 |
| YSCEF1A.PE1 | ELONGATION FACTOR-1-ALPHA 1 | 459 |
| YSCEF1AA.PE1 | EF-1-ALPHA 2 | 459 |
| YSCENOA.PE1 | ENOLASE A | 438 |
| YSCENOB.PE1 | ENOLASE B | 438 |
| YSCG3PDA.PE1 | GLYCERALDEHIDE 3-PHOSPHATE DH-A (G3PD) | 333 |
| YSCG3PDB.PE1 | GLYCERALDEHYDE-3-PHOSPHATE DH-B (G3PD) | 333 |
| YSCG3PDC.PE1 | GLYCERALDEHYDE-3-PHOSPHATE DH PUTATIVE | 333 |
| YSCH2B1.PE1 | HISTONE H2B-1 | 132 |
| YSCH2B2.PE1 | HISTONE H2B-2 | 132 |
| YSCH34CI.PE1 | (C)HISTONE H3-1 | 137 |
| YSCH34CI.PE2 | HISTONE H4-1 | 104 |
| YSCH34CII.PE1 | (C)HISTONE H3-2 | 104 |
| YSCH34CII.PE2 | HISTONE H4 | 104 |
| YSCHSP90.PE1 | HSP90 HEAT SHOCK PROTEIN | 710 |
| YSCLEU2.PE1 | BETA-ISOPROPYLMALATE DEHYDROGENASE | 369 |
| YSCPHO53.PE1 | REPRESSIBLE ACID PHOSPHATASE (PHO5) | 468 |
| YSCPHO53.PE2 | CONSTITUTIVE ACID PHOSPHATASE (PHO3) | 468 |
| YSCRP29.PE1 | RIBOSOMAL PROTEIN 29 | 156 |
| YSCRP51A.PE1 | RIBOSOMAL PROTEIN 51A, EXON 1-2 | 137 |
| YSCRP51B.PE1 | RIBOSOMAL PROTEIN 51B, EXON 1-2 | 137 |
| YSCRPL17A.PE1 | RIBOSOMAL PROTEIN L17A, EXON 1-2 | 138 |
| YSCRPL25.PE1 | RIBOSOMAL PROTEIN L25, EXON 1-2 | 138 |
| YSCRPL29.PE1 | RIBOSOMAL PROTEIN L29 EXON 1-2 | 150 |
| YSCRPS33.PE1 | RIBOSOMAL PROTEIN S33 | 68 |
| YSGRPL16.PE1 | RIBOSOMAL PROTEIN L16 | 175 |
| YSCACT.PE2 | ACTIN EXON 1-2 | 376 |
| YSCCYC1.PE1 | CYC1 | 22 |
| YSCEF1AB.PE1 | ELONGATION FACTOR 1-ALPHA 3 | 459 |
| YSCGDHM.PE1 | GLUTAMATE DEHYDROGENASE 1 | 455 |
| YSCGDHN.PE1 | GLUTAMATE DEHYDROGENASE 2 | 454 |
| YSCHXK1.PE1 | HEXOKINASE P-I | 486 |
| YSCHXK2.PE1 | HEXOKINASE P-II | 487 |
| YSCMRNP.PE1 | POLY (A)-BINDING PROTEIN 1 | 578 |
| YSCPABPG.PE1 | POLYADENYLATE-BINDING PROTEIN 2 | 578 |
| YSCPEP4.PE1 | ASPARTYL PROTEASE PRECURSOR | 406 |
| YSCPORIN.PE1 | PORIN | 284 |
| YSCRPL46.PE1 | RIBOSOMAL PROTEIN L46, EXON 1-2 | 52 |
| YSCRPS24.PE1 | RIBOSOMAL PROTEIN S24 | 131 |
| YSCSUC2.PE1 | PRE-INVERTASE (SECRETED FORM) | 533 |
| YSCSUC2.PE2 | INVERTASE (INTRACELLULAR FORM) | 513 |
| YSCTHS1.PE1 | THREONYL-tRNA SYNTHETASE | 735 |
| YSCUB1G | (s.CEREVISIAE) UBIQUITIN GENE | 80 |
| YSCUB2G3E.PE1 | POLYUBIQUITIN PRECURSOR (aa 39 AT 1) | 192 |
| YSGH4.PE1 | HISTONE H4 (S.CALBERGENSIS) | 104 |
| YSCRPL34.PE1 | RIBOSOMAL PROTEIN L34 | 113 |
| YSCRP13.PE1 | RIBOSOMAL PROTEIN L3 (TCM1) | 387 |
| YSCRP28.PE1 | RIBOSOMAL PROTEIN 28 | 186 |
| YSCRPS10 | RIBOSOMAL PROTEIN S10-2 | 236 |
| YSCRPS16A1 | RIBOSOMAL PROTEIN S16A | 144 |
| YSCH2A1 | HISTONE H2A-1 | 132 |
| YSCH2A2 | HISTONE H2A-2 | 132 |
| YSCPYK | PYRUVATE KINASE | 499 |
| YSCPGK | 3-PHOSPHOGLYCERATE KINASE - PGK GENE | 416 |
| YSCTPI | TRIOSE PHOSPHATE ISOMERASE - TPI GENE | 248 |
| YSCRPS31 | RIBOSOMAL PROTEIN S31 - PUTATIVE | 108 |

different species. There are several possible explanations for these observations, including the tendency to use codons corresponding to most abundant tRNA genes (6,7,8), the theoretical advantage of intermediate bond strengths between tRNA and mRNA (9), the variable total G+C content (10) and the Markov drift in codon usage towards an equilibrium distribution (11). Accordingly a correct codon-usage table may be established by singling out the homogeneous set of coding sequences within which our unknown sequence would most likely be found. In fact cluster analysis method allows for recognition of groups of sequences with the same codon strategy. Obviously, realistic clusterization is possible only after calculation has been made of a correct distance value between the sequences. The distance formula proposed by Grantham (3) does not take into account the influence of aminoacid composition of the

Tab. 2 List of the 99 lowly expressed yeast genes clustered in the second group extracted from the merged GenBank (release 50) and EMBL (release 10) databases.

| locus name | description | codons |
|---|---|---|
| YSCADE4.PE1 | AMIDOPHOSPHORIBOSYLTRANSFERASE | 511 |
| YSCARG4.PE1 | ARGININOSUCCINATE LYASE (EC 4.3.2.1) | 464 |
| YSCCBP1.PE1 | CBP1 PROTEIN | 655 |
| YSCCDC28.PE1 | CDC28 GENE PROTEIN | 299 |
| YSCCDC28A.PE1 | PROTEIN KINASE (CDC28) | 299 |
| YSCCDC8.PE1 | CDC8 GENE | 217 |
| YSCCEN3.PE1 | UNIDENTIFIED READING FRAME | 53 |
| YSCCPA1.PE1 | CARBAMYL PHOSPHATE SYNTHETASE | 412 |
| YSCCPA2.PE1 | CPA2 | 1119 |
| YSCCS.PE1 | CITRATE SYNTHASE | 481 |
| YSCCUP1.PE1 | COPPER CHELATIN | 62 |
| YSCCYC.PE1 | CYTOCHROME C1 PRECURSOR | 310 |
| YSCCYC17.PE1 | CYTOCHROME C REDUCTASE SUBUNIT VI | 148 |
| YSCCYC7.PE1 | ISO-2-CYTOCHROME C | 114 |
| YSCCYCR.PE1 | CYTOCHROME C REDUCTASE 17 KD SUBUNIT | 128 |
| YSCGAL.PE1 | (C)GALACTOKINASE (GAL10) | 46 |
| YSCGAL.PE2 | GALACTOKINASE (GAL1) | 29 |
| YSCGAL1P.PE1 | GALACTOSE-1-PHOSPHATE URIDYLYL TRANSFERASE | 21 |
| YSCGAL4.PE1 | GAL4 PROTEIN | 882 |
| YSCGCN4.PE1 | GCN4 PROTEIN | 282 |
| YSCGCN4B.PE1 | GENERAL CONTROL PROTEIN (GCN4) | 250 |
| YSCHIS4.PE1 | HIS4 POLYPEPTIDE | 800 |
| YSCHMLAL.PE1 | (C)PROTEIN ALPHA-2 A | 211 |
| YSCHMLAL.PE2 | PROTEIN ALPHA-1 A | 176 |
| YSCLEU1.PE1 | ISOPROPYLMALATE-1 ISOMERASE | 48 |
| YSCM1P1.PE1 | M1-P1 PREPROTOXIN | 317 |
| YSCM1PPT.PE1 | PREPROTOXIN | 317 |
| YSCMATA.PE1 | PROTEIN A1 EXON 1-2-3 | 127 |
| YSCMATAL.PE1 | (C)PROTEIN ALPHA-2 B | 212 |
| YSCMATAL.PE2 | PROTEIN ALPHA-1 | 176 |
| YSCODCD.PE1 | OROTIDINE-5'-PHOSPHATE DECARBOXYLASE MONOMER-1 | 268 |
| YSCODCF.PE1 | OROTIDINE-5'-PHOSPHATE DECARBOXYLASE MONOMER-2 | 268 |
| YSCPPR2.PE1 | PYRIMIDINE PATHWAY REGULATORY 2 PROTEIN | 129 |
| YSCPUT2.PE1 | DELTA-1-PYRROLINE-5-CARBOXYLATE DEHYDROGENASE | 576 |
| YSCRAD1.PE1 | RAD1 PROTEIN (PUTATIVE) | 973 |
| YSCRAS1 | RAS1 GENE, COMPLETE CODING SEQUENCE | 310 |
| YSCRAS2 | RAS2 GENE, COMPLETE CODING SEQUENCE | 323 |
| YSCRASH1R.PE1 | RAS PROTEIN 1 | 310 |
| YSCRASH2R.PE1 | RAS PROTEIN 2 | 323 |
| YSCTKCDC8.PE1 | THYMIDYLATE KINASE | 217 |
| YSCTRP1.PE1 | TRP1 (N-(5'-PHOSPHORIBOSYL)-ANTHRANILATE ISOM. | 225 |
| YSCTRP2.PE1 | ANTHRANILATE SYNTHASE COMPONENT I | 529 |
| YSCTRP3.PE1 | ANTHRANILATE SYNTHASE COMPONENT II (TRP3) | 485 |
| YSCTUBB.PE1 | BETA-TUBULIN | 458 |
| YSCYP2ONC.PE2 | YP2 PROTEIN | 207 |
| YSDSTA1.PE1 | PREPROGLUCOAMYLASE | 779 |
| YSGGALS1.PE1 | (C)GAL7 | 185 |
| YSGGALS1.PE2 | (C)GAL10-1 | 399 |
| YSGGALS2.PE1 | (C)GAL10-2 | 46 |
| YSGGALS2.PE2 | GAL1 | 529 |
| YSCH3P | PARTIAL HISTONE 3 (H3) | 34 |
| YSCADE.PE1 | GLYCINEAMIDE RIBOTIDE SYNTASE/ARS POLYPROT. | 803 |
| YSCADH3.PE1 | ALCOHOL DEHYDROGENASE III | 376 |
| YSCARG3.PE1 | ORNITHINE CARBAMOYLTRANSFERASE (EC 2.1.3.3) | 339 |
| YSCCAR.PE1 | ARGINASE | 334 |
| YSCCDC7.PE1 | CDC7 GENE PRODUCT | 508 |
| YSCCPAX.PE1 | CARBAMOYL-PHOSPHATE SYNTHETASE SMALL SUBUNIT | 412 |
| YSCCYC1X.PE1 | ISO-1-CYTOCHROME C | 110 |
| YSCCYC4.PE1 | CYTOCHROME C OXIDASE SUBUNIT IV PREPEPTIDE | 156 |
| YSCCYCPX | NUCL. MIT. CYTOCHROME C PEROXIDASE GENE. | 80 |
| YSCGAL7.PE1 | GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE | 366 |
| YSCGAL80G.PE1 | GAL80 REGULATORY PROTEIN | 436 |
| YSCLEU4.PE1 | ALPHA-ISOPROPYLMALATE SYNTHASE L (EC 4.1.3.12) | 620 |
| YSCLEU4.PE2 | ALPHA-ISOPROPYLMALATE SYNTHASE S (EC 4.1.3.12) | 590 |
| YSCMEL1.PE1 | ALPHA-GALACTOSIDASE PRECURSOR | 472 |
| YSCMFA1G.PE1 | ALPHA-FACTOR-1 PHEROMONE PRECURSOR | 166 |
| YSCMFA2G.PE1 | ALPHA-FACTOR-2 PHEROMONE PRECURSOR | 121 |
| YSCMSS51.PE1 | MSS51 PROTEIN | 437 |
| YSCMSW.PE1 | TRYPTOPHANYL-Trna SYNTHETASE | 375 |
| YSCPET9.PE1 | ADP/ATP TRANSLOCATOR | 310 |
| YSCPK25.PE1 | PROTEIN KINASE | 398 |
| YSCPLASM.PE1 | (C)REP 1 PROTEIN | 374 |
| YSCPLASM.PE2 | D PROTEIN | 182 |
| YSCPLASM.PE3 | (C)REP 2 PROTEIN | 297 |
| YSCPLASM.PE4 | RECOMBINASE (FLP) | 424 |
| YSCPPR1.PE1 | REGULATORY PROTEIN | 905 |
| YSCRAD2G.PE1 | RAD2 PROTEIN | 1032 |
| YSCRAD3.PE1 | RAD3 PROTEIN 1 | 779 |
| YSCRAD3G.PE1 | RAD3 PROTEIN 2 | 779 |
| YSCRAD6.PE1 | RAD6 PROTEIN | 173 |
| YSCRPO21.PE1 | RNA POLYMERASE II LARGE SUBUNIT | 1727 |
| YSCRPO31.PE1 | RNA POLYMERASE III LARGE SUBUNIT | 1461 |
| YSCSIR2G.PE1 | SIR2 PROTEIN | 563 |
| YSCSIR3G.PE1 | SIR3 PROTEIN | 979 |
| YSCSPT2.PE1 | SPT2 PROTEIN | 334 |
| YSCSTE2G.PE1 | STE2 PROTEIN | 432 |
| YSCSTE3.PE1 | PHEROMONE A RECEPTOR | 471 |
| YSCSTE3G.PE1 | STE3 PROTEIN PRECURSOR | 471 |
| YSCSTE6PR.PE1 | STE6 PROTEIN | 41 |
| YSCSTE7.PE1 | STE7 PROTEIN | 516 |
| YSCSUC2A.PE1 | PRE-INVERTASE (SECRETED FORM) | 30 |
| YSCSUC7.PE1 | INVERTASE PREPEPTIDE | 74 |
| YSCTOP2.PE1 | TOPOISOMERASE II | 1430 |
| YSCTOPI.PE1 | TOPOISOMERASE I | 770 |
| YSCURA3.PE1 | ORF | 29 |
| YSGMAL6ST.PE1 | (C)MALTOSE PERMEASE (MAL6T) | 31 |
| YSGMAL6ST.PE2 | MALTASE | 585 |
| YSGMEL1.PE1 | PRE-ALPHA GALACTOSIDASE (MELIBIASE) | 472 |
| YSCMTRGF.PE1 | R1 TRANSLATION PRODUCT | 236 |

Tab. 3 Codon usage table of the 59 highly expressed yeast genes, reporting total number of occurrences, percent frequency, codon usage value and standard deviation for each codon (total codon numbers : 17237).

| Am.acid | Codon | Number | Freq % | Cod-use | Sd.dev. |
|---|---|---|---|---|---|
| Arg | CGA | 0 | 0.00 | 0.00 +/- | 0.00 |
| Arg | CGC | 4 | 0.02 | 0.00 +/- | 0.02 |
| Arg | CGG | 0 | 0.00 | 0.00 +/- | 0.00 |
| Arg | CGT | 129 | 0.75 | 0.15 +/- | 0.13 |
| Arg | AGA | 690 | 4.00 | 0.82 +/- | 0.14 |
| Arg | AGG | 22 | 0.13 | 0.03 +/- | 0.05 |
| Leu | CTA | 103 | 0.60 | 0.08 +/- | 0.08 |
| Leu | CTC | 5 | 0.03 | 0.00 +/- | 0.01 |
| Leu | CTG | 33 | 0.19 | 0.02 +/- | 0.03 |
| Leu | CTT | 40 | 0.23 | 0.03 +/- | 0.04 |
| Leu | TTA | 218 | 1.26 | 0.16 +/- | 0.10 |
| Leu | TTG | 934 | 5.42 | 0.70 +/- | 0.16 |
| Ser | TCA | 84 | 0.49 | 0.07 +/- | 0.07 |
| Ser | TCC | 372 | 2.16 | 0.32 +/- | 0.13 |
| Ser | TCG | 16 | 0.09 | 0.01 +/- | 0.02 |
| Ser | TCT | 561 | 3.25 | 0.49 +/- | 0.13 |
| Ser | AGC | 48 | 0.28 | 0.04 +/- | 0.05 |
| Ser | AGT | 69 | 0.40 | 0.06 +/- | 0.08 |
| Thr | ACA | 66 | 0.38 | 0.07 +/- | 0.08 |
| Thr | ACC | 394 | 2.29 | 0.40 +/- | 0.15 |
| Thr | ACG | 15 | 0.09 | 0.02 +/- | 0.02 |
| Thr | ACT | 519 | 3.01 | 0.52 +/- | 0.13 |
| Pro | CCA | 578 | 3.35 | 0.81 +/- | 0.16 |
| Pro | CCC | 17 | 0.10 | 0.02 +/- | 0.05 |
| Pro | CCG | 9 | 0.05 | 0.01 +/- | 0.03 |
| Pro | CCT | 107 | 0.62 | 0.15 +/- | 0.14 |
| Ala | GCA | 74 | 0.43 | 0.05 +/- | 0.06 |
| Ala | GCC | 440 | 2.55 | 0.31 +/- | 0.12 |
| Ala | GCG | 18 | 0.10 | 0.01 +/- | 0.03 |
| Ala | GCT | 908 | 5.27 | 0.63 +/- | 0.17 |
| Gly | GGA | 38 | 0.22 | 0.03 +/- | 0.04 |
| Gly | GGC | 79 | 0.46 | 0.06 +/- | 0.09 |
| Gly | GGG | 25 | 0.15 | 0.02 +/- | 0.03 |
| Gly | GGT | 1200 | 6.96 | 0.89 +/- | 0.14 |
| Val | GTA | 29 | 0.17 | 0.02 +/- | 0.04 |
| Val | GTC | 521 | 3.02 | 0.41 +/- | 0.11 |
| Val | GTG | 58 | 0.34 | 0.05 +/- | 0.06 |
| Val | GTT | 654 | 3.79 | 0.52 +/- | 0.12 |
| Lys | AAA | 324 | 1.88 | 0.23 +/- | 0.15 |
| Lys | AAG | 1115 | 6.47 | 0.77 +/- | 0.15 |
| Asn | AAC | 592 | 3.43 | 0.80 +/- | 0.15 |
| Asn | AAT | 149 | 0.86 | 0.20 +/- | 0.15 |
| Gln | CAA | 582 | 3.38 | 0.96 +/- | 0.07 |
| Gln | CAG | 23 | 0.13 | 0.04 +/- | 0.07 |
| His | CAC | 203 | 1.18 | 0.64 +/- | 0.27 |
| His | CAT | 112 | 0.65 | 0.36 +/- | 0.27 |
| Glu | GAA | 994 | 5.77 | 0.89 +/- | 0.10 |
| Glu | GAG | 123 | 0.71 | 0.11 +/- | 0.10 |
| Asp | GAC | 510 | 2.96 | 0.53 +/- | 0.16 |
| Asp | GAT | 456 | 2.65 | 0.47 +/- | 0.16 |
| Tyr | TAC | 449 | 2.60 | 0.78 +/- | 0.18 |
| Tyr | TAT | 125 | 0.73 | 0.22 +/- | 0.18 |
| Cys | TGC | 19 | 0.11 | 0.14 +/- | 0.16 |
| Cys | TGT | 117 | 0.68 | 0.86 +/- | 0.16 |
| Phe | TTC | 496 | 2.88 | 0.71 +/- | 0.19 |
| Phe | TTT | 204 | 1.18 | 0.29 +/- | 0.19 |
| Ile | ATC | 474 | 2.75 | 0.48 +/- | 0.15 |
| Ile | ATT | 496 | 2.88 | 0.50 +/- | 0.13 |
| Ile | ATA | 23 | 0.13 | 0.02 +/- | 0.04 |
| Met | ATG | 348 | 2.02 | 1.00 +/- | 0.00 |
| Trp | TGG | 170 | 0.99 | 1.00 +/- | 0.00 |
| *** | TGA | 5 | 0.03 | 0.11 +/- | 0.31 |
| *** | TAA | 34 | 0.20 | 0.76 +/- | 0.43 |
| *** | TAG | 6 | 0.03 | 0.13 +/- | 0.34 |

Tab. 4 Codon usage table of the 99 lowly expressed yeast genes, reporting total number of occurrences, percent frequency, codon usage value and standard deviation for each codon ( total codon number: 39745).

| Am.acid | Codon | Number | Freq % | Cod-use | Sd.dev. |
|---|---|---|---|---|---|
| Arg | CGA | 112 | 0.28 | 0.06 +/- | 0.07 |
| Arg | CGC | 77 | 0.19 | 0.04 +/- | 0.05 |
| Arg | CGG | 63 | 0.16 | 0.03 +/- | 0.05 |
| Arg | CGT | 358 | 0.90 | 0.20 +/- | 0.13 |
| Arg | AGA | 884 | 2.22 | 0.48 +/- | 0.15 |
| Arg | AGG | 333 | 0.84 | 0.18 +/- | 0.10 |
| Leu | CTA | 507 | 1.28 | 0.14 +/- | 0.06 |
| Leu | CTC | 202 | 0.51 | 0.06 +/- | 0.04 |
| Leu | CTG | 400 | 1.01 | 0.11 +/- | 0.06 |
| Leu | CTT | 453 | 1.14 | 0.13 +/- | 0.06 |
| Leu | TTA | 977 | 2.46 | 0.27 +/- | 0.10 |
| Leu | TTG | 1051 | 2.64 | 0.29 +/- | 0.10 |
| Ser | TCA | 648 | 1.63 | 0.21 +/- | 0.09 |
| Ser | TCC | 474 | 1.19 | 0.15 +/- | 0.07 |
| Ser | TCG | 290 | 0.73 | 0.09 +/- | 0.05 |
| Ser | TCT | 956 | 2.41 | 0.31 +/- | 0.10 |
| Ser | AGC | 302 | 0.76 | 0.10 +/- | 0.07 |
| Ser | AGT | 445 | 1.12 | 0.14 +/- | 0.08 |
| Thr | ACA | 733 | 1.84 | 0.32 +/- | 0.11 |
| Thr | ACC | 488 | 1.23 | 0.21 +/- | 0.10 |
| Thr | ACG | 294 | 0.74 | 0.13 +/- | 0.07 |
| Thr | ACT | 758 | 1.91 | 0.33 +/- | 0.11 |
| Pro | CCA | 772 | 1.94 | 0.44 +/- | 0.13 |
| Pro | CCC | 266 | 0.67 | 0.15 +/- | 0.10 |
| Pro | CCG | 192 | 0.48 | 0.11 +/- | 0.09 |
| Pro | CCT | 538 | 1.35 | 0.30 +/- | 0.12 |
| Ala | GCA | 688 | 1.73 | 0.28 +/- | 0.10 |
| Ala | GCC | 582 | 1.46 | 0.24 +/- | 0.10 |
| Ala | GCG | 248 | 0.62 | 0.10 +/- | 0.07 |
| Ala | GCT | 918 | 2.31 | 0.38 +/- | 0.12 |
| Gly | GGA | 401 | 1.01 | 0.17 +/- | 0.12 |
| Gly | GGC | 425 | 1.07 | 0.18 +/- | 0.09 |
| Gly | GGG | 248 | 0.62 | 0.11 +/- | 0.08 |
| Gly | GGT | 1260 | 3.17 | 0.54 +/- | 0.19 |
| Val | GTA | 494 | 1.24 | 0.20 +/- | 0.11 |
| Val | GTC | 529 | 1.33 | 0.21 +/- | 0.09 |
| Val | GTG | 458 | 1.15 | 0.18 +/- | 0.09 |
| Val | GTT | 1004 | 2.53 | 0.40 +/- | 0.13 |
| Lys | AAA | 1651 | 4.15 | 0.57 +/- | 0.13 |
| Lys | AAG | 1222 | 3.07 | 0.43 +/- | 0.13 |
| Asn | AAC | 873 | 2.20 | 0.41 +/- | 0.13 |
| Asn | AAT | 1275 | 3.21 | 0.59 +/- | 0.13 |
| Gln | CAA | 1024 | 2.58 | 0.69 +/- | 0.16 |
| Gln | CAG | 459 | 1.15 | 0.31 +/- | 0.16 |
| His | CAC | 283 | 0.71 | 0.36 +/- | 0.17 |
| His | CAT | 509 | 1.28 | 0.64 +/- | 0.17 |
| Glu | GAA | 1861 | 4.68 | 0.71 +/- | 0.10 |
| Glu | GAG | 764 | 1.92 | 0.29 +/- | 0.10 |
| Asp | GAC | 809 | 2.04 | 0.34 +/- | 0.11 |
| Asp | GAT | 1577 | 3.97 | 0.66 +/- | 0.11 |
| Tyr | TAC | 670 | 1.69 | 0.46 +/- | 0.14 |
| Tyr | TAT | 786 | 1.98 | 0.54 +/- | 0.14 |
| Cys | TGC | 180 | 0.45 | 0.33 +/- | 0.18 |
| Cys | TGT | 368 | 0.93 | 0.67 +/- | 0.18 |
| Phe | TTC | 667 | 1.68 | 0.40 +/- | 0.14 |
| Phe | TTT | 987 | 2.48 | 0.60 +/- | 0.14 |
| Ile | ATC | 633 | 1.59 | 0.25 +/- | 0.10 |
| Ile | ATT | 1275 | 3.21 | 0.51 +/- | 0.12 |
| Ile | ATA | 612 | 1.54 | 0.24 +/- | 0.14 |
| Met | ATG | 853 | 2.15 | 1.00 +/- | 0.00 |
| Trp | TGG | 442 | 1.11 | 1.00 +/- | 0.00 |
| *** | TGA | 51 | 0.13 | 0.38 +/- | 0.38 |
| *** | TAA | 56 | 0.14 | 0.41 +/- | 0.41 |
| *** | TAG | 29 | 0.07 | 0.21 +/- | 0.33 |

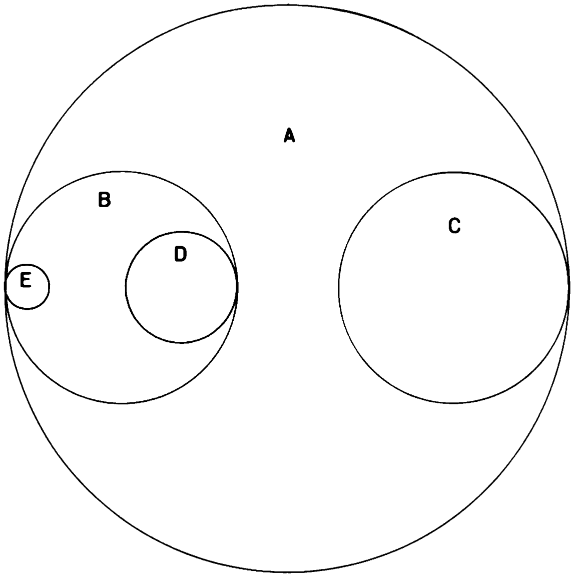


Fig. 2 Density map relative to 158 yeast genes, allowing for selection immediate of subclusters in the main cluster. Circle diameters correspond to the maximum distance between genes in each circle.

different gene products, while those of Gribskov (12) and Sharp (5) do not consider the total occurrences of the synonymous codons for each family and thus do not allow for the contribution of the codon usage of each family in calculating the distance between sequences.

Rather than being based on the clustering procedure, the improvement of our method compared to the others lies in a more correct evaluation of the distance between sequences. In fact we propose a chi2-like distance formula as

Tab. 5 Density and circle diameter for each cluster. Diameter and density computed respectively according to the formulas [5] and [6].

| Cluster | N. of genes | diameter | density |
|---|---|---|---|
| A | 158 | 10.99 | 0.23 |
| B | 59 | 4.42 | 1.30 |
| C | 99 | 4.50 | 2.07 |
| D | 47 | 2.14 | 9.16 |
| E | 12 | 0.78 | 48.29 |

| GENE | N. CODONS | N. DEGENERATE NUCLEOTIDES | PERCENT IDENTITY | | | |
|---|---|---|---|---|---|---|
| | | | Y | LY | HY | R |
| Enolase B | 349 | 544 | 55.1 | 48.0 | 69.7 | 41.4 |
| Ribosomal protein L3 | 387 | 470 | 52.6 | 45.7 | 69.8 | 38.5 |
| Poly(A) - binding protein 1 | 578 | 692 | 55.8 | 52.3 | 63.7 | 43.2 |
| ATPase beta subunit | 313 | 396 | 53.2 | 51.0 | 59.1 | 38.4 |
| Repressible acid phosphatase (PHO 5) | 468 | 576 | 50.3 | 45.8 | 58.3 | 38.7 |
| Constitutive acid phoasphatase (PHO 3) | 468 | 563 | 47.1 | 41.4 | 54.2 | 35.2 |
| Ribosomal protein S33 | 68 | 87 | 42.5 | 35.6 | 54.1 | 33.3 |
| Pre-invertase | 533 | 653 | 49.8 | 46.7 | 54.4 | 42.6 |

Tab. 6 Percent identity for 8 highly expressed yeast gene sequences obtained by backtranslating protein sequences using (Y) codon usage table computed on all 158 yeast genes, (HY) codon usage table computed on the 59 highly expressed genes, (LY) codon usage table computed on the 99 lowly expressed genes and (R) randomly with no codon usage table. Percent identity was evaluated between backtranslated and real sequences by considering only degenerate codon positions.

the basis for cluster analysis which substantially improves the above mentioned formulas, as confirmed by data obtained when our method is applied to yeast genes.

Furthermore, sequences backtranslated using our method contain fewer errors than either a merely random sequence or one backtranslated according to a codon-usage table obtained without cluster analysis.

On the basis of the codon usage table, our method chooses the position of the

Tab. 7 Results of the most likely probe selection performed on 7 highly expressed yeast genes.

| GENE | PROBE LENGTH | OPTIMAL PROBE POSITION (aa) | | N. OF MATCHES |
|---|---|---|---|---|
| | | From | to | |
| Enolase B | 36 | 57 | 68 | 35 |
| Ribosomal protein L3 | 36 | 113 | 124 | 35 |
| Poly(A) - binding protein 1 | 36 | 456 | 467 | 35 |
| Alcohol dehydrogenase II | 36 | 98 | 109 | 35 |
| Ribosomal protein 51A | 36 | 73 | 84 | 36 |
| Ribosomal protein L29 | 36 | 73 | 84 | 34 |
| Actin | 36 | 305 | 316 | 35 |

optimal probe in the polipeptide sequence as that containing the fewest ambiguous aminoacids. Moreover it may produce for this region any number of randomly weighed probe sequences, among which further selection may be carried out on the basis of such criteria as dinucleotide frequencies and G-T base pairing in order to obtain a more unified optimal probe (13).

The codon-usage table generated by our method could also be used for optimal expression of heterologous genes in DNA-recombinant biotechnological applications using hosts such as E.coli and yeast, whose codon bias is closely linked to gene expression (14,15).

ACKNOWLEDGEMENTS

We are grateful to Prof. G. Preparata for his assistance in constructing the model and curating stimulating discussions,to Prof. C. Saccone for his many helpful observations and to Prof. John Credico for preparing the final English version.

This work was partially financed by the M.P.I. (Italy) through the "Progetto Finalizzato Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie" (CNR) and by the "Progetto Strategico delle Biotecnologie" (CNR) - Italy.

## REFERENCES

1. Devereux, J., Haeberli, P., Smithies, O. (1984) Nucl. Acids Res. 12(1), pp.387-395.
2. Lewis, M.L. (1986) Nucl. Acids Res. 14(1), pp.567-570.
3. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R. (1981) Nucl. Acids Res. 9(1), pp.143-174.
4. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and Di Paola G. (1985) CABIOS 1(3), pp.167-172.
5. Sharp, M.P., Touhy, T.M.F. and Mosursky, K.R. (1986 ) Nucl. Acids Res. 14(13), pp.5125-5143.
6. Ikemura, T. (1981) J. Mol. Biol. 146, pp.1-21.
7. Ikemura, T. (1981) J. Mol. Biol. 151, pp.389-409.
8. Ikemura, T. (1982) J. Mol. Biol. 158, pp.573-579.
9. Grosjean, H. and Fiers, W. (1982) Gene 18, pp.199-209.
10. Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24, pp.1-11.
11. Wilbur, W.J. (1985) J. Mol.Evol . 21, pp.169-181.
12. Gribskov, M., Devereux, J. and Burgess, R.R. (1984) Nucl. Acids Res. 12(1), pp.539-549.
13. Lather, R. (1985) J. Mol. Biol. 183, 1-12.
14. Gouy, M. and Gautier, C. (1982) Nucl. Acids. Res. 10, 7055-7074.
15. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.