Evolution of Sequences, Structures and Genomes

Protein and Peptide Science Group/Industrial Biochemistry and Biotechnology Group Colloquium Organized and Edited by D. Higgins (Department of Biochemistry, University College Cork). 670th Meeting, held at University College Cork, 7–9 September 1999.

Evidence in favour of ancient octaploidy in the vertebrate genome

T. J. Gibson*[1] and J. Spring†
*European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany, and †Institute of Zoology, University of Basel, Rheinsprung 9, CH-4051 Basel, Switzerland

Abstract

Vertebrate genomes are larger than invertebrates and show evidence of extensive gene duplication, including many collinear chromosomal segments. On the basis of this intra-genomic synteny, it has been proposed that two rounds of whole genome duplication (octaploidy) occurred early in the vertebrate lineage. Recently, this early vertebrate octaploidy has been challenged on the basis of gene trees. We report new linkage groups encompassing the matrilin (MATN), syndecan (SDC), Eyes Absent (EYA), HCK kinase and SRC kinase paralogous gene quartets. In contrast to other studies, the sequence trees are weakly supportive of ancient octaploidy. It is concluded that there is no strong evidence against the octaploidy, provided that consecutive genome duplication was rapid.

Introduction

Some years ago, Ohno [1] suggested that polyploidy had been an important factor in the evolution of vertebrates, genetic redundancy providing the material for adaptive divergence. These conjectures were based primarily upon genome content, chromosome topologies and the high frequency of tetraploid fish and amphibians. Ohno [1] elaborated his proposals before the confounding effects of pre-mRNA splicing and other forms of non-coding DNA could have been anticipated.

In the last few years, as the extent of genetic redundancy throughout the vertebrates has become clear, a number of authors have revisited these ideas, leading to a proposition that there were two rounds of genome duplication (octaploidy) early in the vertebrate lineage [2–9]. Thus there are four *HOX* gene clusters in vertebrates but only one in *Drosophila*. Many other gene families show up to four paralogous genes in the vertebrates for each orthologous invertebrate gene. The precise timing of the two genome duplication events is not yet clear, but the earliest date would be about 500 million years ago (Mya). The single *HOX* gene cluster in the Chordate *Amphioxus* suggests that the duplications occurred after this lineage had split from proto-vertebrates [10], while PCR studies suggest that the jawless lamprey has at least three *HOX* clusters [11,12]. However, nothing is known about the *HOX* genes (or any other informative multigene families) for the early diverged hagfish, another jawless vertebrate. The timing uncertainty would allow either (1) a rapid octaploidy occurring in a few hundred or thousand years, or (2) tetraploidy followed by a process of diploidization over several million years (Myrs) and then a second tetraploidy.

Recently, the validity of the proposed early vertebrate octaploidy has been questioned. Skrabanek and Wolfe [13] have argued that the available data from the human genome are currently insufficient to test the model. Hughes [14] has studied a number of multigene families, concluding that both the number of vertebrate paralogues (with respect to the closest invertebrate gene) and

Key words: evolution, gene duplication, gene trees, paralogous genes, tetraploidy.

Abbreviations used: DDC, duplication–degeneration–complementation model; EYA, Eyes Absent; MATN, matrilin; Mya, million years ago; Myrs, million years; SDC, syndecan.

[1]To whom correspondence should be addressed.

the gene tree topologies are in conflict with the model.

It is of some importance to resolve whether or not the ancient vertebrate polyploidy is correct. Quadruplication of vertebrate genes immediately provides an explanation for the extensive genetic redundancy observed in the vertebrates [15]. It would also provide a fixed time point for comparative studies of the evolution of gene families, in particular to measure the variation in mutation rates within and between multigene families. As has been pointed out earlier [15,16], current evolutionary models would not adequately explain the persistence of redundant genes for hundreds of Myrs: this should provide a stimulus to evolutionary theorists.

Vertebrate genomes show extensive intragenomic synteny, that is the presence of collinear regions on different chromosomes that are known to be duplicate, triplicate or quadruplicate. It is these syntenic regions that have led to the proposal of vertebrate octaploidy. Gene trees used as evidence against the polyploidy [14] have not been made from the gene sets that provide the strongest evidence in favour—those that are found within extensive syntenies. This is important as the only way to control for the consistency of the trees is to use linked gene sets. We therefore set out to review the data and to search for sets of syntenic gene quartets that could be used to assess whether tree topologies would be informative.

Materials and methods

Linked gene families were collected using a combination of Internet resources, such as Locus Link (http://www.ncbi.nlm.nih.gov/LocusLink/), OMIM (http://www.ncbi.nlm.nih.gov/omim/) and MGI (http://mgd.hgmp.mrc.ac.uk/) together with sequence homology searches for paralogues. Amino acid sequences were aligned in Clustal X [17] and corrected by hand as needed. Neighbour-Joining trees [18] were then calculated, excluding unreliable and gapped regions and correcting for multiple substitutions.

Results and discussion

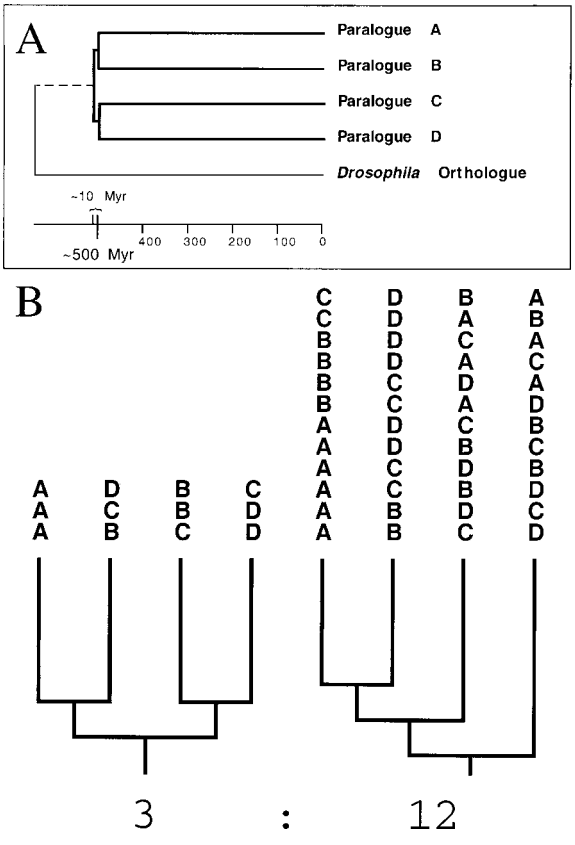
When should gene trees be informative?

The generally expected tree for a true octaploid groups the four paralogues into two pairs (Figure 1A). In the vertebrate case, the time for the accumulation of informative substitutions before the second genome duplication would have been short with respect to the time available to accumulate gene-specific substitutions, i.e. the sequences might have difficulty in resolving the true tree, even if 10 Myrs or more elapsed between consecutive tetraploidies. In fact, the period between duplications could have been much less than 10 Myrs. Many examples of rapid octaploid formation are known from crop plants under domestication and recent octaploids are found in the frog species *Ceratophrys dorsata* and *C. ornata* [19,20]. *C. ornata* is quite remarkable: this 'species' actually consists of diploid, tetraploid and octaploid populations [20]. In both species, the octaploid chromosomes form octavalents at meiosis, so that all eight alleles at each locus are freely exchangeable. This means that any attempt to resolve the history of the frog genome quadruplication by constructing gene trees would be wholly meaningless. These modern examples show that we cannot be certain whether gene trees

Figure 1

Trees for four related genes

(**A**) Phylogenetic tree expected for a vertebrate paralogous gene family arising 500 Mya as a consequence of two genome duplications separated by 10 Myrs. Internal branches are very short compared with terminal branches. (**B**) All possible rooted tree topologies for four related genes. Only three of the 15 topologies are compatible with octaploidy.



can test the model for ancient vertebrate octaploidy.

Whenever sequence trees for gene quartets lack informative data for resolving internal branches, the tree topology will be essentially random. As shown in Figure 1 (B), there are four times more tree topologies that are inconsistent with the expected octaploid tree. Therefore, since we cannot be certain that the tree topologies will be informative, we need to examine trees from linked gene families so that we can have confidence in the shared history of the gene duplications. In this way, we can control for the stability and consistency of the tree topologies using the neighbouring genes. The gene families examined by Hughes [14] were unlinked, leading to uncertainties in their histories of duplication and deletion [6].

Syntenic quartets within the human genome

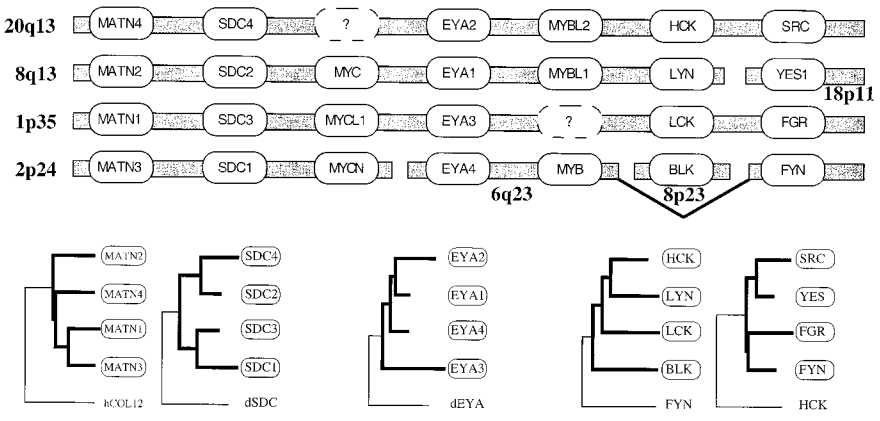
Three major groups of quadruplicated collinear chromosomal regions have been presented in the literature as evidence for ancient vertebrate octaploidy. The *HOX* gene clusters, with linked collagen and epidermal growth factor receptor gene quartets reside on chromosomes 2, 7, 12 and 17. They provide at least 15 paralogous gene sets, only five of which, however, still retain all four paralogues [2,21]. Since the *HOX* homeodomains are so highly conserved, and silent mutations are likely to be saturated [22], they cannot be used to test the order of duplication of the clusters. The MHC-related regions are found on chromosomes 1, 6 (MHC), 9 and 19 [23,24]. Of the 11 paralogous gene sets, only the notch-like receptors and the tumour necrosis factor-like cytokines are known to retain all four paralogues [4]. The third published syntenic quartet links portions of chromosomes 4, 5, 8 and 10 [8]. It includes quartets of fibroblast growth factor and adrenergic receptors, but interpretation is complicated by earlier tandem duplications of the adrenergic receptors and later chromosomal rearrangements. Therefore none of the published gene sets is currently ideal for providing sequence tree tests.

We searched genome maps and recent publications for gene families that might reveal additional collinear quartets and found two groupings of note. First, we found two linked gene quartets for the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) glutamate receptors and the androgen/mineralocorticoid/glucocorticoid/progesterone nuclear receptor family on chromosomes X, 4, 5 and 11. Since there are only two paralogous gene quartets, it would be premature to further analyse this set. However, the recent cloning and mapping of a fourth human Eyes Absent (EYA) paralogue [25] allowed us to build up linkage groups centred on chromosomes 1, 2, 8 and 20 that include gene quartets for the matrilins (MATN), syndecans (SDC) and the eight SRC/HCK-related tyrosine kinases, albeit with the need to invoke chromosomal rearrangements disrupting the chromosome 2 and 8 linkage groups (Figure 2). While these rearrangements limit confidence in the analysis, they

Figure 2

The linked MATN, SDC, EYA, HCK and SRC gene quartets

Linkage of seven paralogous gene families and their approximate chromosomal locations. Exact gene order is mostly unknown. Only three paralogues are known for the *MYC* and *MYB* families. An earlier tandem duplication underlies the HCK and SRC quartets. Two translocations and an inversion are posited following the octaploidy. Neighbour-Joining trees for each quartet (thick branches) are rooted by the closest available outgroup sequence (thin branches). Horizontal branch lengths are proportional to divergence. The best supported tree topology (SDC and SRC) is compatible with octaploidy.



mainly affect the highly investigated kinases, for which no new paralogues have been discovered for a number of years, nor are any new additions expected. This linkage set is the only one we could find that provided a reasonable number of linked gene quartets (five) for which we could compare sequence trees.

Comparison of sequence trees from a set of five syntenic gene quartets

Figure 2 shows Neighbour-Joining trees for each protein quartet, rooted by the closest available outgroup sequence. The five trees yielded four different topologies. Bootstrapping (not shown) tests indicated that none of the topologies was stable, consistent with the short internal branch lengths. The SDC and the SRC quartets gave the same tree topology, one that is consistent with an octaploidy. The MATN, EYA and HCK quartets each gave a different topology that would be inconsistent with an octaploidy: however, in each case one pair of sequences (MATN1/MATN3; EYA1/EYA3; HCK/LYN) was grouped as in the SDC/SRC tree topology. The latter topology could be recreated for the MATN tree if the MATN2 branch moved through one node. Moving the EYA3 and BLK branches through one node would similarly recover the SDC/SRC topology. The EYA and HCK topologies can also be reconciled by moving one branch. However, the MATN topology would require two sequence branches to be moved to align to the EYA and HCK trees. Therefore the best supported topology is clearly the SDC/SRC tree, which is consistent with an octaploidy.

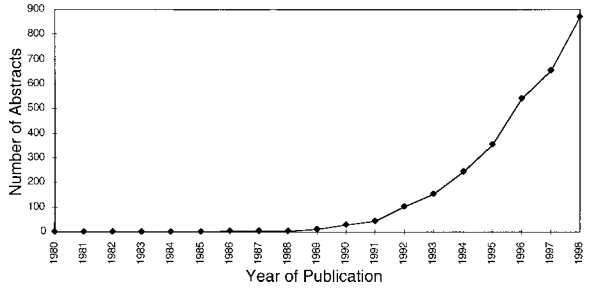
The EYA3 sequence is the deepest-branched vertebrate EYA paralogue. By reference to the linked MATN1 and SDC3 genes, the EYA3 sequence has undergone accelerated evolution with respect to its three paralogues (Figure 2). Most probably EYA3 is under less purifying selection than EYA1, 2 and 4. Since the EYA tree violates the assumption of a molecular clock—a common cause of artefactual trees—EYA3 can be assumed to be incorrectly placed, most probably as a consequence of long-branch attraction. To be consistent with the linked MATN1 and SDC3, EYA3 should be grouped with EYA4, which would yield the SRC/SDC tree.

Overall, the five trees provide some support for a topology consistent with an ancient octaploidy. This is nevertheless a weak result, owing to the small number and instability of the trees as well as the posited chromosomal rearrangements. More data are needed.

Figure 3

Growth of 'dominant negative' usage in journal abstracts

The annual usage of 'dominant negative' was retrieved from the PubMed server (http://www.ncbi.nlm.nih.gov/PubMed/). Before 1988, the concept was barely discussed in molecular and medical genetics. Advances in methodology have revealed that dominant negative phenotypes are common for multidomain proteins and in many rare human inherited conditions, such as Marfan's syndrome.



Persistence of duplicate genes

Modern molecular genetics, mouse gene knockouts in particular, have revealed extensive functional redundancy in regulatory and cytoskeletal proteins, such as signalling proteins and transcription factors. By contrast, enzymes of intermediary metabolism show little redundancy and usually have single-copy genes [15,16,26].

Force et al. [27] have recently elaborated a theory [the duplication–degeneration–complementation (DDC) model] for the preservation of duplicate genes through complementary, degenerative mutations in gene promoters. The effect of these mutations would be the partitioning of ancestral functions, in particular the neutral evolution of complementary tissue specificities. This theory is consistent with a large body of evidence on the tissue-specific expression of vertebrate paralogous gene sets, such as the SRC-like kinases. However, the theory does not seem to explain the differential preservation of redundant gene classes.

Evolutionary models of duplicate gene fates have usually not distinguished between genes with recessive and dominant phenotypes [15,16]. It seems likely that the significance of dominant mutations has been underestimated in the past, probably because severe phenotypes are embryonic lethals and easily overlooked, while *de novo* point mutations are untraceable. There are three major classes of dominant lesions: (1) haploid insufficiency, usually because a set of closely

interacting proteins require 'balanced gene dosage'; (2) loss of heterozygosity, as in tumour suppressors, where a somatic mutation in the good allele promotes cancer; and (3) dominant negatives, which are almost always domain-specific lesions in multidomain proteins. Figure 3 shows that the literature on dominant negative phenotypes is only 10 years old, yet more than 1000 papers are expected to discuss dominant negative phenotypes in 1999. Many of these refer to progress in understanding rare dominantly inherited human genetic diseases, but most discuss dominant negative experimental constructs.

In the vertebrates, the most redundant gene families belong to classes that often show dominant phenotypes, whereas recessively inherited metabolic enzymes overwhelmingly possess single-copy genes. We have argued that redundant, dominantly inherited genes will persist in the genome much longer than recessive genes, since most deleterious point mutations will be rejected by purifying selection [15]. This will allow time for neutral, complementary promoter mutations to alter the gene expression patterns according to the DDC model [27], ultimately fixing the genes as they become increasingly tissue-specific.

The combination of 500 Myr-old vertebrate octaploidy, dominant inheritance and the DDC model may account for the observed differential retention of redundant gene classes. However, until the vertebrate octaploidy can be confirmed, there remains a formal possibility that differential duplication, not differential retention, is the cause of this disparity.

Conclusions

The evidence for an ancient vertebrate octaploidy rests upon the burgeoning quantity of collinear chromosomal segments currently found in pairs, triples and quadruples, but not apparently in higher numbers. The currently mapped genes are insufficient to confirm the model. Although gene trees have been held as evidence against the octaploidy [14] because they often do not show the expected topology, they cannot be used to assess the evolutionary history of an octavalent octaploid. The sequence trees can be taken as evidence against models with a long period of diploidization between genome duplications or models of allo-octaploidy (interspecific hybrization) [4], unless the parental species were so close that octavalents resulted. The MATN, SDC, EYA, HCK and SRC trees presented here are weakly consistent with octaploidy.

We conclude, therefore, that there is still no strong evidence against the ancestral octaploidy, but that the second round of genome duplication must have followed rapidly upon the first. Ultimately, vertebrate octaploidy seems likely to be confirmed (or otherwise) by the human genome sequencing project, which will enable the full genome to be assessed for the degree of fourfold duplication. If the synteny is found to be very extensive, it should be sufficient, regardless of gene tree topologies.

We thank José Castresana for discussions.

References

1 Ohno, S. (1970) Evolution by Gene Duplication, Springer-Verlag, Berlin
2 Holland, P. W., Garcia-Fernandez, J., Williams, N. A. and Sidow, A. (1994) Dev. Suppl. 125–133
3 Holland, P. W. and Garcia-Fernandez, J. (1996) Dev. Biol. **173**, 382–395
4 Spring, J. (1997) FEBS Lett. **400**, 2–8
5 Lundin, L. G. (1993) Genomics **16**, 1–19
6 Smith, N. G., Knight, R. and Hurst, L. D. (1999) Bioessays **21**, 697–703
7 Nadeau, J. H. and Sankoff, D. (1997) Genetics **147**, 1259–1266
8 Pébusque, M. J., Coulier, F., Birnbaum, D. and Pontarotti, P. (1998) Mol. Biol. Evol. **15**, 1145–1159
9 Sidow, A. (1996) Curr. Opin. Genet. Dev. **6**, 715–722
10 Garcia-Fernandez, J. and Holland, P. W. (1996) Int. J. Dev. Biol. Suppl. **40**, 71S–72S
11 Sharman , A. C. and Holland, P. W. (1998) Int. J. Dev. Biol. **42**, 617–620
12 Pendleton, J. W., Nagai, B. K., Murtha, M. T. and Ruddle, F. H. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 6300–6304
13 Skrabanek, L. and Wolfe, K. H. (1998) Curr. Opin. Genet. Dev. **8**, 694–700
14 Hughes, A. L. (1999) J. Mol. Evol. **48**, 565–576
15 Gibson, T. J. and Spring, J. (1998) Trends Genet. **14**, 46–49; discussion 49–50
16 Cooke, J., Nowak, M. A., Boerlijst, M. and Maynard-Smith, J. (1997) Trends Genet. **13**, 360–364
17 Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) Nucleic Acids Res. **25**, 4876–4882
18 Saitou, N. and Nei, M. (1987) Mol. Biol. Evol. **4**, 406–425
19 Beçak, M. L., Beçak, W. and Rebello, M. N. (1967) Chromosoma **22**, 192–201
20 Schmid, M., Haff, T. and Schempp, W. (1985) Chromosoma **91**, 172–184
21 Bailey, W. J., Kim, J., Wagner, G. P. and Ruddle, F. H. (1997) Mol. Biol. Evol. **14**, 843–853
22 Zhang, J. and Nei, M. (1996) Genetics **142**, 295–303
23 Katsanis, N., Fitzgibbon, J. and Fisher, E. M. C. (1996) Genomics **35**, 101–108
24 Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T. and Ishibashi, T. (1996) Proc. Natl. Acad. Sci. U.S.A. **93**, 9096–9101

25 Borsani, G., DeGrandi, A. Ballabio, A., Bulfone, A., Bernard, L., Banfi, S., Gattuso, C., Mariani, M., Dixon, M., Donnai, D. et al. (1999) Hum. Mol. Genet. **8**, 11–23
26 Iwabe, N., Kuma, K. and Miyata, T. (1996) Mol. Biol. Evol. **13**, 483–493
27 Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999) Genetics **151**, 1531–1545

Received 9 September 1999

Searching for the ideal forms of proteins

W. R. Taylor
Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.

Abstract

A modification of the Structure Alignment Program (SAP), combined with a novel automatic method for the definition of structural elements, correctly identified the core folds of a variety of small β/α proteins when compared with a series of ideal architectures. This approach opens the possibility of not just determining whether one structure is like another, but given a range of ideal forms, determining what the protein is. Preliminary studies have shown it to work equally well on the all α-class and the all-β class of protein, each of which have corresponding ideal forms. Given the speed of the algorithm, it will be possible to compare all of these against the Protein Structure Database and determine the extent to which the current ideal forms can account for the variety of protein structure. Analysis of the remainder should provide a base for the development of further forms.

Introduction

With the large number of protein structures now known, it is difficult to gain an overview of their variety of forms and even more difficult to comprehend how each structure relates to its neighbours. Systematic attempts have been made to instil order into this bewildering variety, most notably in the heirarch classifications captured in the SCOP [1], CATH [2] and DALI [3] structure databases. The order in these collections is based on the pairwise comparison of protein structures using either intuitive (SCOP), automatic (DALI) or combined (CATH) approaches. The classification of structures based on comparison is strongest when the proteins are most similar, so all these collections differ little in their allocation of similar structures. The more difficult task is when there is some similarity in the fold of more than two proteins with each having different features in common. In this situation, the decision to group proteins together can often be arbitrary or, more cautiously, not made. The latter solution leads to a large number of distinct entities and rather than producing a tall classification tree, results in a low bush.

This situation is similar to that of the naturalist of the 19th Century who classified groups based on numbers of legs, teeth, bones and other features, giving rise to strong relationships between similar animals (or plants using different features of course) but less as the similarity became more distant, and if the organisms shared no common features, then they could not be considered related. To move on from this 'collecting' phase requires the adoption of an underlying theory that can structure the different groups given only weak evidence and, in the absence of data, can provide a default relationship as a working model until proven otherwise. For the naturalist, the underlying theory was provided through the ideas of evolution and ultimately through the modern phylogenetic analysis of sequence data. It might seem that a definitive resolution of the protein classification problem could also be attained along similar lines. However, protein structures are more strongly conserved through evolution than the sequences that embody them, which implies that the difficult areas in protein structure classification cannot be resolved through sequence comparison.

Without recourse to an evolutionary history, an alternative approach is to search for unifying structural principles that can be represented as idealized protein—protein archetypes, or their underlying Platonic forms. An indication of what

Key words: protein architecture, protein classification, protein comparison, protein structure.
Abbreviations used: SAP, Structure Alignment Program; RMS, root mean square; PDB, Protein Structure Database; 3chy, chemotaxis-Y protein; 5 nul, short-chain flavodoxin; 2fcr, long-chain flavodoxin; 1 etu, ribosomal elongation factor Tu; 5p21, ras oncogene protein p21; 3adk, adenylate kinase; 1kev, alcohol dehydrogenase.