ELSEVIER

FEMS Microbiology Reviews 22 (1998) 207–227

FEMS MICROBIOLOGY REVIEWS

Review

Indigo: a World-Wide-Web review of genomes and gene functions

P. Nitschké [a,*], P. Guerdoux-Jamet [b], H. Chiapello [c], G. Faroux [a], C. Hénaut [d], A. Hénaut [a], A. Danchin [d]

[a] Université de Versailles Saint Quentin, Laboratoire Génome et Informatique, 45 avenue des États Unis, 78000 Versailles, France
[b] UPRES-A 6026, Équipe Canaux et Récepteurs Membranaires, Campus de Beaulieu, 35042 Rennes Cedex, France
[c] INRA Laboratoire de Biologie Cellulaire, route de St Cyr, 78026 Versailles Cedex, France
[d] Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Received 15 May 1998; received in revised form 19 July 1998; accepted 22 August 1998

Abstract

The present article describes a genome database reviewing gene-related knowledge of two model bacteria, *Bacillus subtilis* and *Escherichia coli*. The database, Indigo, is open through the World-Wide Web (http://indigo.genetique.uvsq.fr). The concept used for organising the data, the concept of neighbourhood, allows one to explore the database content in an efficient although somewhat unusual way. Here, genes are related to each other by a variety of neighbourhoods, including proximity in the chromosome, phylogenetic kinship, participation in a common metabolic pathway, common presence in an article of the literature, or similar use of the genetic code. Several examples illustrate how this concept of neighbourhood permits one to review the available knowledge about a given gene or gene family, and elaborate unexpected, but revealing, analyses about gene functions. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: Database; Genomics; Neighborhood; Factorial correspondence analysis; tRNA synthetase

Contents

[*] Corresponding author. Tel.: +33 (1) 3925-4559; Fax: +33 (1) 3925-4569; E-mail: patrick.nitschke@genetique.uvsq.fr

PII: S0168-6445(98)00014-X

1. Introduction

Genome sequences can be retrieved from a small number of computer centres. However, the biological knowledge associated with the deciphering of most genome texts usually results from the collaboration of scientists all over the world. It has therefore become necessary to organise the corresponding vast amount of knowledge as genome sequences are generated (at the time of this review, one complete bacterial genome each month). To fully benefit from this information, we need to put together the disparate available pieces of information related to a gene and its function(s). For a long time, this has been the main goal of review articles, that provided the readers with integrated and up to date information on a given topic. In parallel, for many years, international organisations maintained and spread data libraries for DNA and protein sequences, together with the software needed to manage and extract the data deposited at their sites. At present, an international consortium provides most of the DNA sequences produced in the world (at the time when they are validated by their authors) as a single DNA library with three entry points: USA (Genbank [1]); Europe (EMBL/EBI [2]); and Japan (DDBJ [3]). The best annotated protein data library, Swiss-Prot, is maintained at the University of Geneva. It is completed by an automatically generated protein complement of the GenBank/EBI/DDBJ DNA library, TREMBL [4].

In addition to these general repositories, specialised data bases for microbial genomes started to appear some years ago with annotations specifically included to improve the comprehensive knowledge about specific organisms. These databases allow extraction of aggregate information for a given organism, as the sequence is collected in ongoing genome projects, or when the complete genome is known (examples can be found at http sites such as: www.pseudomonas.com, www.qmw.ac.uk/~rhbm001/salmopage.html, chlamydia-www.berkeley.edu:4231, genome-www.stanford.edu/Saccharomyces [5], tigr.org, and in particular at a site that collects information on most on-going projects, www-c.mcs.anl.gov/home/gaasterl/genomes.html [6]). Model genome data bases started to be constructed ten years ago. For example, Kröger and coworkers [7] have collected the *Escherichia coli* sequences from EBI/Genbank (http://susi.bio.uni-giessen.de/ecdc.html) since 1989. Similar data bases were developed in Japan ([8], http://mol.genes.nig.ac.jp/ecoli, http://genome4.aist-nara.ac.jp). At the same time, Médigue et al. [9] proposed a relational database for the *E. coli* genome: Colibri, initially managed by the DataBase Management System (DBMS) 4th Dimension, and now by the Sybase DBMS (http://www.pasteur.fr/Bio/Colibri.html). This work has been exploited for constructing relational databases for other genomes, such as *Saccharomyces cerevisiae* [10], or *Bacillus subtilis* (SubtiList http://www.pasteur.fr/Bio/SubtiList.html [11]). Related to these sequence databases, paper references in scientific magazines, reviews or books have been organised as well, allowing extraction of bibliographical data. In the medical domain, a literature library, Medline, was developed 20 years ago for this purpose by the National Library of Medicine. It now contains citations from more than 9 millions articles in the biomedical field. To search Medline, the NCBI (National Center for Biotechnology Information) conceived and constructed a WWW browser, Entrez [12], that allows retrieval of bibliographical citations integrated in MedLine (http://www4.ncbi.nlm.nih.gov/PubMed). Among other interesting features (in particular the feature of ‘neighbourhood’, that we shall discuss below), links have been established between data contained in sequences database and in related bibliographical information. Finally, several servers present metabolic databases, related to the present knowledge of genome data. The best examples derive from the work of Karp et al. (EcoCyc: [13]), Selkov and Overbeek and coworkers (PUMA, WIT: [14,15]), and Kanehisa and coworkers (KEGG: [16]).

A popular, but laborious, means to review the knowledge about a given organism is to write a series of review articles, such as the remarkable compendium published by the American Society for Microbiology on *E. coli* and *S. typhimurium* [17]. However, this requires an extensive effort that can only be renewed every decade or so. Computer-mediated approaches can tackle the problem of regular updating much more smoothly and efficiently. We present such an approach, the construction of a World-Wide Web server, Indigo (http://indigo.genetique.uvsq.fr), devoted to specific genomes, using the concept of gene neighbourhood. Indigo creates an interactive environment allowing the interested scientists to retrieve and exploit the knowledge about gene neighbours for model organisms (at present: *E. coli* and *B. subtilis*, and a preliminary compendium of *A. thaliana* genes). We review here how the concept of gene neighbourhood, intrinsically associated to the many links a gene or its product(s) can have with other biological entities can help computer-mediated study of genomes (in silico analysis, for short [18]).

2. Presentation

2.1. Gene neighbourhoods

The main idea underlying the present review is that the biological objects making a cell alive cannot be isolated from each other: biology must be described more as a science of relationships between objects, than as a science describing objects. In short, a living organism cannot be summarised as the collection of its genes and gene products. Knowledge of whole genome sequences is a unique opportunity to study the relationships between genes and gene products at the level of the cell, the atom of heredity. In most cases, we ignore what relationships are involved, but we know that they exist. To study them, we investigated the concept of neighbourhood in order to organise the disparate knowledge we have on model bacteria, *E. coli* and *B. subtilis*. This concept is very wide: it pertains to the category of notions that John Myhill named 'prospective characters' [19]. It can, however, be visualised as making, in a broad sense, reference to all the items, of all possible kinds, that can be related to a given item. Because we study the genomic text, we chose *genes* as the core items. For a given gene, we constructed lists of neighbours based on links of several possible categories. This provides us with a new type of review of the information content of genomes as a whole.

The first and intuitive relationship between two genes is their proximity in the chromosome. Two genes are considered as geographical neighbours if they are in close vicinity in the chromosome. Although the concepts of operon, or in a broader sense, of pathogenicity islands, are clearly related to such proximity, this kind of relationship is far from sufficient to explain *functional* relationships between genes. We therefore included in our database many other kinds of gene neighbourhoods. As an example, a second possibility, often used in classical studies [12], is to relate genes or gene products because they evolve from a common ancestor. This constitutes families of paralogues (inside a given genome), or orthologues (between different genomes). Paralogous or orthologous genes constitute clusters of genes with similar sequences. Of course, there is no reasons to reduce possible neighbourhoods to these two kinds of relationships. For example, we can consider that genes coding for proteins involved in the same metabolic pathways, or using the same substrate are related. This constitutes the metabolic neighbourhoods. More complex relationships have been described, such as relationships based on the genetic code utilisation or on common presence in bibliographical references: two genes can be related because they used synonymous codons with the same frequency; they also can be linked because they are cited in the same bibliographical source. At Indigo, the following neighbourhoods have been taken into account: operons, alignments with entries in Swiss-Prot, metabolic pathways, codon usage, similarity in isoelectric points of the gene products and literature. This list is not exhaustive and in the future the site will evolve, using other kinds of neighbourhoods.

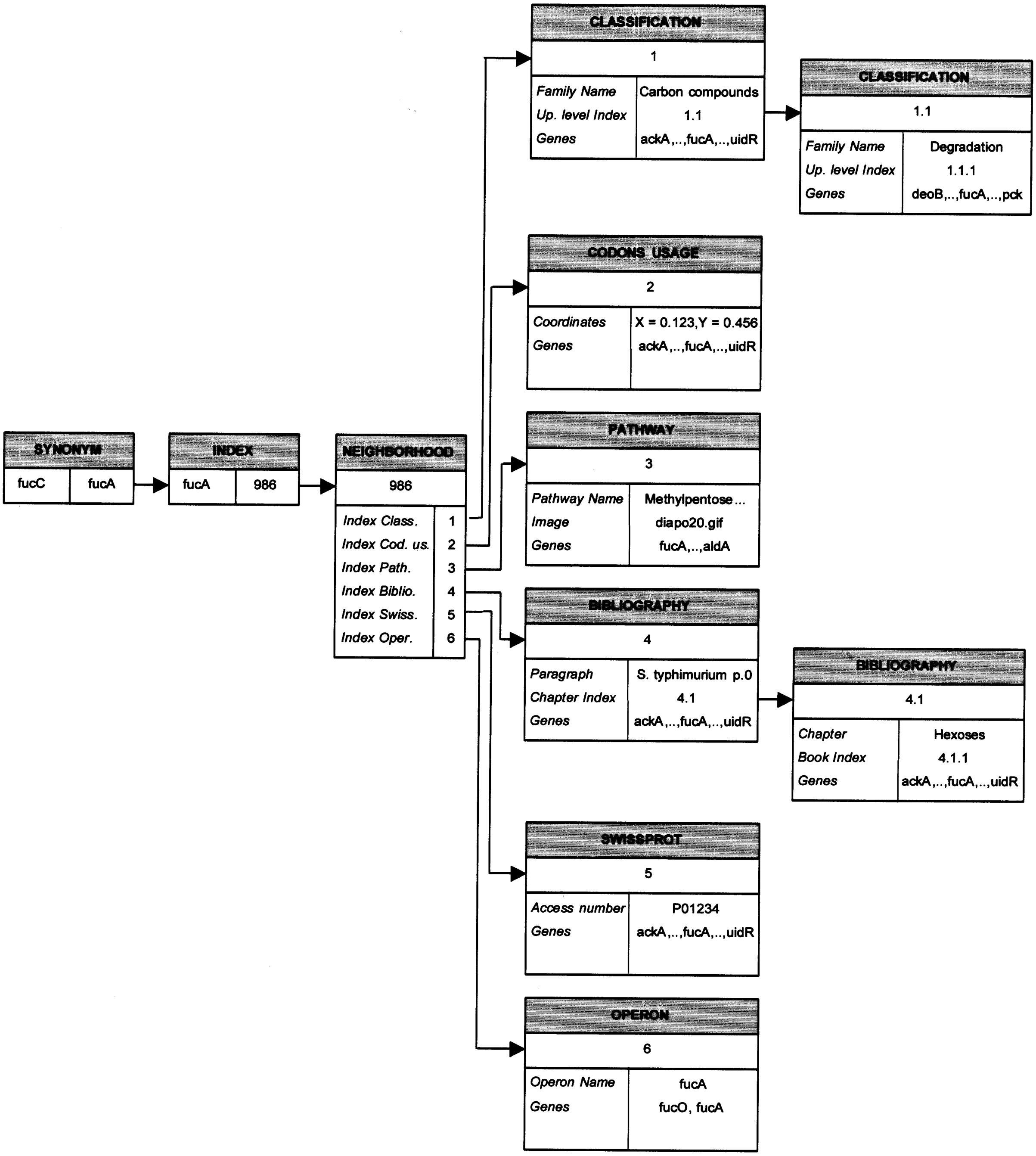


Fig. 1. The Indigo database structure. The http server (indigo.genetique.uvsq.fr) is organised around neighbourhoods of all genes of model organisms. At present, four main neighbourhoods are used: functional classification, codon usage, metabolic pathways and literature.

2.2. Software development

The structure underlying Indigo was written using the JAVA language (Sun Microsystems) and compiled with the JAVA Development Kit 1.0.2. We chose to develop Indigo in JAVA because this language has two major advantages over other languages: compatibility with most computers and operating systems, and both a network-oriented use (via the WWW) or a stand-alone use (e.g. using CD-ROMs). The version described in this work (version 1.0.2) is compatible with Netscape Navigator 3.0 (Netscape) and Internet Explorer 3.0 (Microsoft) or higher.

JAVA is, however, a very demanding language, that requires powerful computers to operate smoothly. We have therefore developed in parallel a standard hypertext markup language (HTML) version (cgi-bin version) of Indigo. As a matter of fact, this 'light' version of Indigo offers less interactivity for the user (it uses only the capabilities of HTML), but it uses and produces the same information as does the full JAVA version. Finally, in order to allow a stand alone use of Indigo, we wanted to remain independent of commercial database management systems. In parallel to being incorporated into Indigo, all data have therefore been stored in an indexed text format (flat files).

2.3. Datasets

At the time of this article, Indigo succinctly reviews our knowledge of two bacterial species, *E. coli* and *B. subtilis*. It manages six different categories of data: Swiss-Prot files, functional classification, operon classification, bibliographical neighbourhoods, codon usage, paralogue families and metabolic pathways (extracted from the literature).

Bibliographical data have been extracted manually from reference monographs and selected review articles (bibliographical neighbourhoods as well as images of biosynthetic pathways and/or of protein complexes, cell architecture or regulatory networks). At this point unless otherwise stated, literature references have not been included directly in the server because of copyright obligations. For any given gene, they can be recovered using PubMed, searching for the name of the gene, and that of each of the genes present in its literature neighbours. The other data have been automatically extracted from specialised servers: Swiss-Prot (http://expasy.hcuge.ch/sprot/sprot-top.html, Swiss-Prot files), operon classification (KEGG: http://www.genome.ad.jp/kegg/kegg2.html), and functional classification (KEGG, GenprotEC: http://www.mbl.edu/html/ecoli.html, and SubtiList: http://www.pasteur.fr/Bio/SubtiList.html). Sequence data for codon usage and paralogue families were extracted for *E. coli* from http://www.ncbi.nlm.nih.gov (accession number U00096) and for *B. subtilis* from the SubtiList database (Data release R14.2).

As stated above, the principle underlying the neighbourhood concept is to gather data from a variety of sources (servers, bibliographies, sequence databases), that are heterogeneous in nature (images, biosynthetic pathways; hierarchies, classifications, etc.). In order to get Indigo as expandable and easily updated as possible we kept its underlying logical structure fairly simple: the relationships between the different pieces of information are based on the only item common to the various neighbourhood's lists: the gene name. This is implemented into the core data of the server as a gene name index. From the selection of a name in this index, access to the different lists of neighbours follows immediately, as shown in Fig. 1. This simplicity allows one to use the Internet, via the World-Wide Web and JAVA applets fairly rapidly, once applets have been transferred to the user's computer.

2.4. General Indigo utilisation

Because an on-line help is integrated into the server we do not develop here in detail all the capabilities of the Indigo interface. The Indigo structure is simple: a first module named 'Selection Module', is the start point of all queries. It gives access to the core of the structure through two information modules, the 'Neighbourhood Module' and the 'Codon Usage Module'. Navigation between these three modules, is as illustrated in Fig. 2.

2.4.1. The selection module

The Selection Module is the first frame displayed when opening Indigo. Selection can be performed using two different approaches, either by typing di-

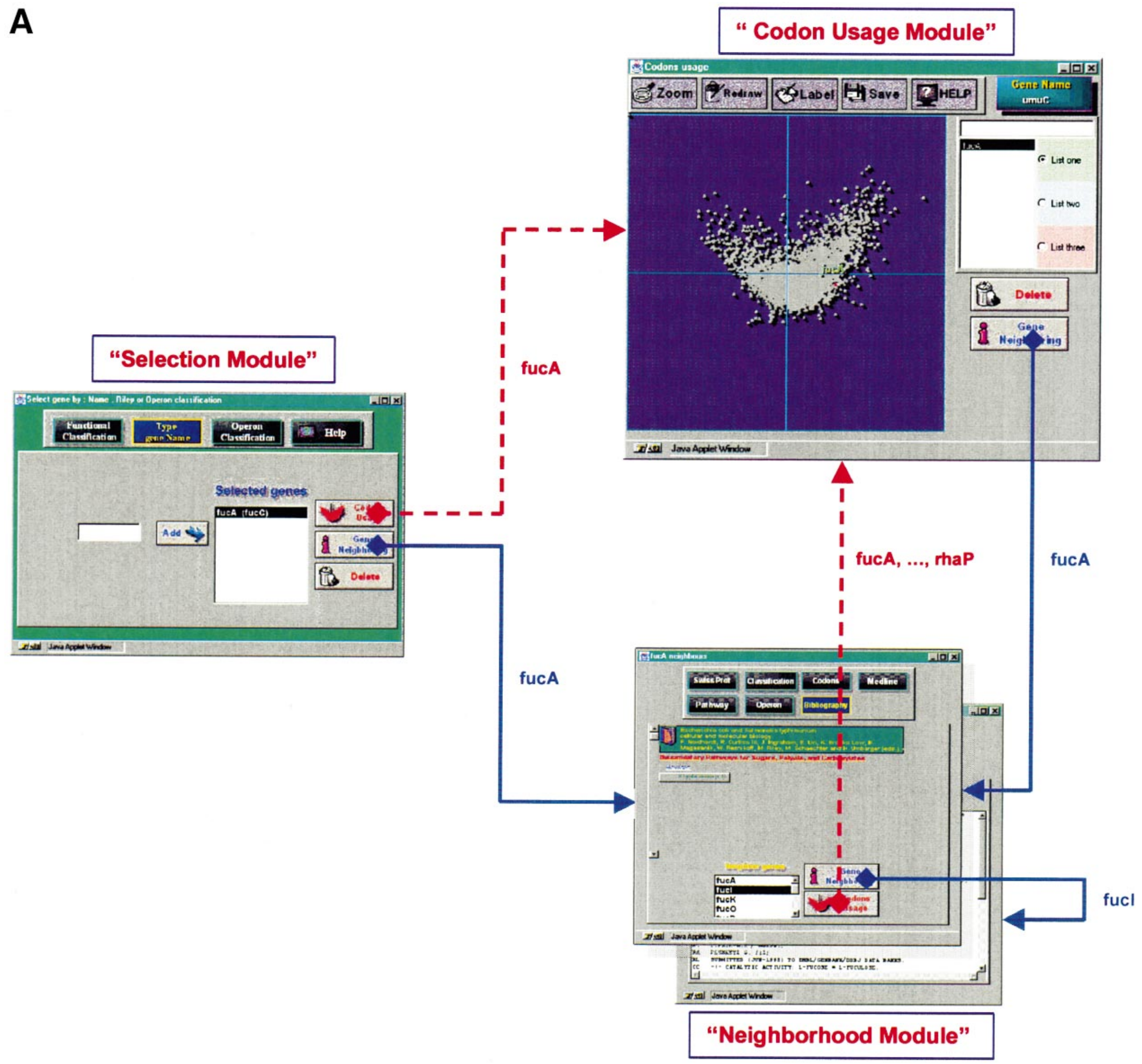


Fig. 2. Navigation in the Indigo server. Starting from one of the three main modules (selection module, codon usage module and neighbourhood information module), one can go to any module either by entering and/or clicking on a gene name, or by clicking on one of the selection buttons associated to a gene name. (A) General organisation. (B) The selection panel. (C) The neighbourhood information module.

rectly a gene name in the appropriate field box, or by searching for a name using an operon or functional classification. Three options can be chosen from the three buttons laid in the upper part of the window. The gene names that have been selected appear in the 'Selected genes' list. The user may subsequently have access to the information present in the database, on either one, several, or all the genes present in the list, by simply selecting the desired genes from the list and clicking on the 'gene neighbourhood' or 'codon usage' buttons.

2.4.2. *The neighbourhood information module*

This module is the main module of Indigo. It gives

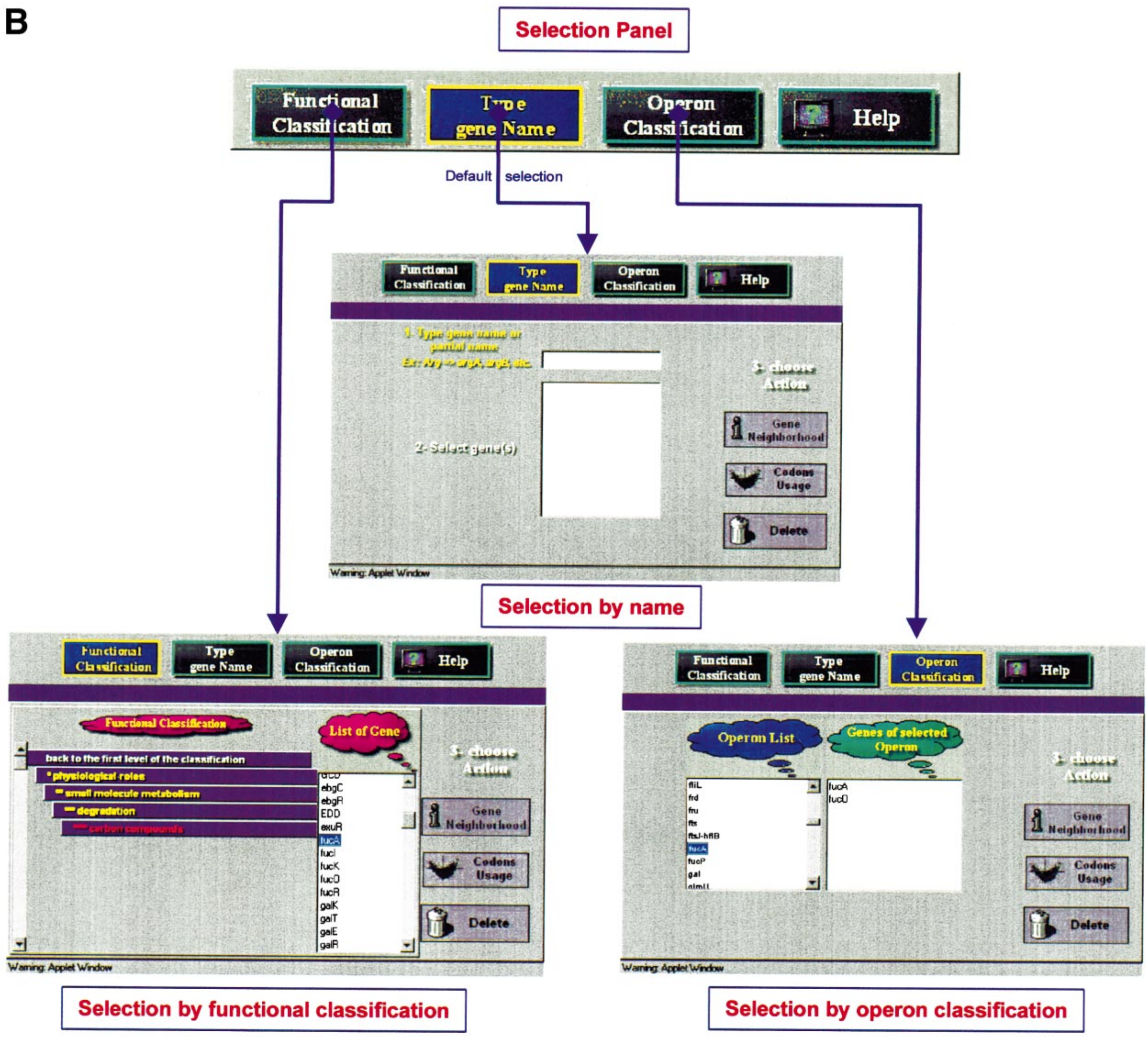
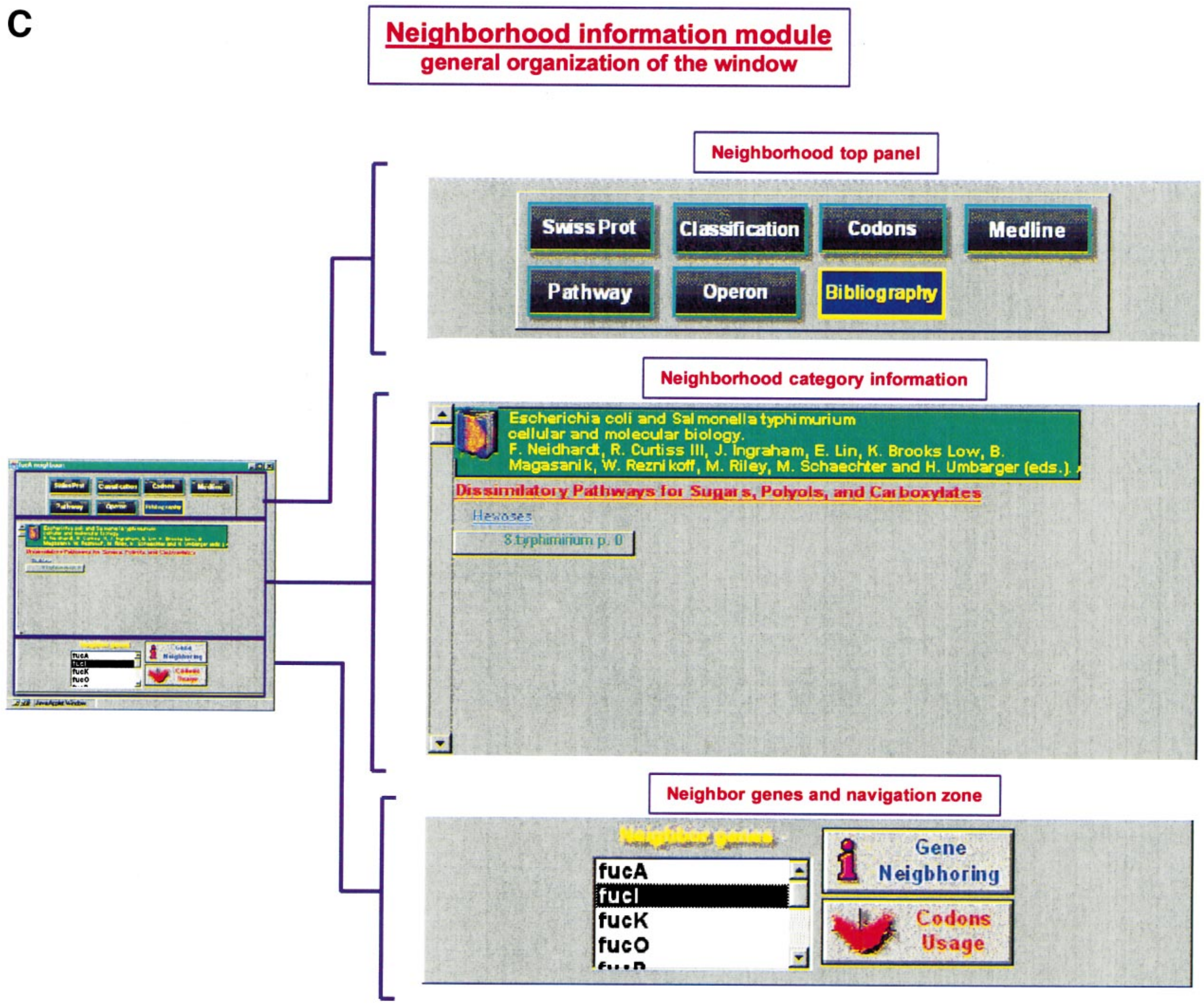


Fig. 2 (*Continued.*)

access to all the neighbourhood data collected in the database. The top panel of the window displays buttons for each neighbour's category available for a given gene (i.e. when one category is not available for the selected gene, the corresponding button is not displayed). When the user selects a category, the window's central panel displays specific information concerning the selected category, and neighbours of the selected gene are shown in the 'neighbour genes' list. If, in a category, the selected gene belongs to several classes (e.g. a gene whose product is involved in several biosynthetic pathways), then the user has to select one of these classes in order to display the corresponding neighbours in the 'neighbour genes' list.

In the case of hierarchical classifications (functional and bibliographical classifications), the entire hierarchical branching containing the selected gene, is displayed to let the user choose the level of neighbourhood he or she wishes to consider (e.g. chapter, sub-chapters and paragraphs are levels in the bibliographical classification).



Escherichia coli and Salmonella typhimurium cellular and molecular biology. F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Brooks Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter and H. Umbarger (eds.).

Dissimilatory Pathways for Sugars, Polyols, and Carboxylates

Hexoses

S.typhimirium p. 0

Fig. 2 (*Continued*.)

2.4.3. The codon usage module

This module provides an interactive representation of the normalised codon usage cloud of all the genes of a given genome, displayed as a two-dimensional projection [20] (Fig. 3). This approach, factorial correspondence analysis (FCA for short, see [21] for a description of the method), derives from the work of Sneath, Sokal and co-workers in statistical taxonomy (see for example [22]). Genes are represented in the 61st-dimensional codon space as a hyperellipsoid with axes displaying decreasing inertia (in fact, as the superimposition of hyperellipsoids of points representing each class of genes, when there are several classes). In the case of codon usage for *E. coli* and *B. subtilis*, most of the inertia is carried by the two first axes. As a consequence, a bi-dimensional representation of the cloud of genes reflects well the proximity of genes according to their codon usage: this justifies the choice of the 2D display of the gene cloud as an opening window for exploring gene neighbourhood according to their codon usage. Through this module, any list of genes obtained with the two previous modules (‘Selected genes’ list from the gene name selection module, or ‘neighbour genes’ list from the neighbourhood information module) can be instantaneously visualised on the codon usage plot. At this

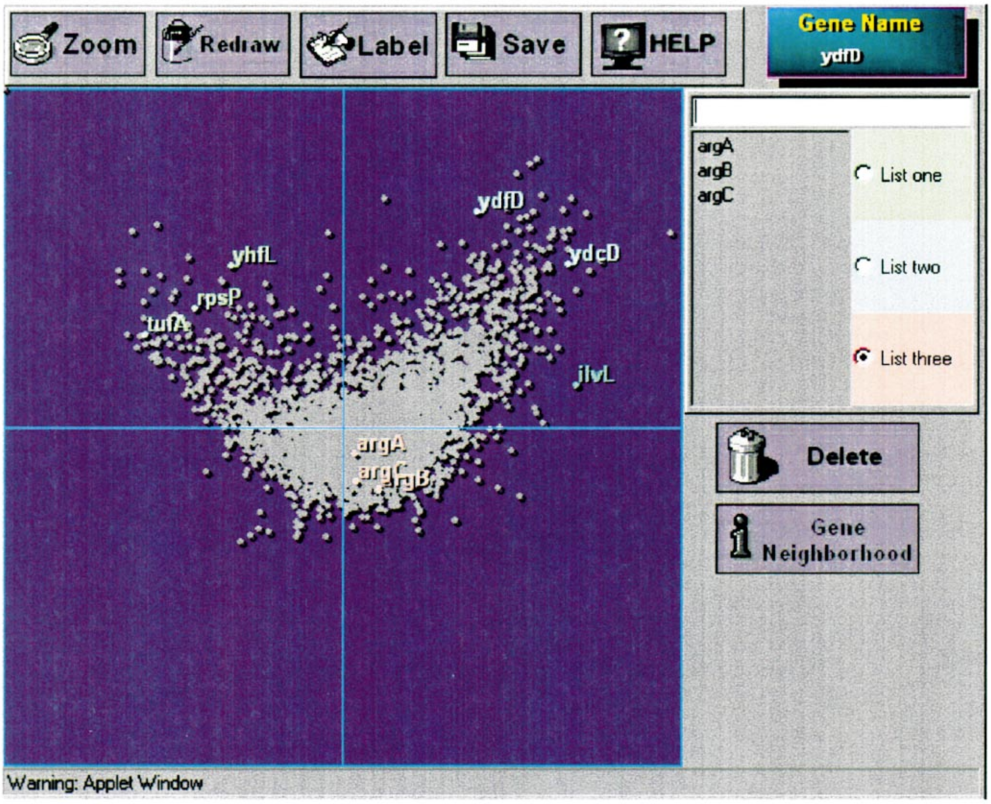


Fig. 3. The codon usage module. Genes can be selected on the spot, and their neighbours can be visualised by clicking and dragging the mouse, then using the zoom button.

point, the user can follow many options, such as doing a close up of an area of the plot, showing gene names of all the points, visualising gene names and selecting them instantaneously from the plot, as well as adding new names to the list or creating a new list.

3. A review of *E. coli* and *B. subtilis*: Indigo

Indigo collects as much of the knowledge as possible about model genomes. In particular, we endeavoured to make it behave as a review of the literature on *E. coli* and *B. subtilis*. The aim of this section is to present a few examples that illustrate the ability to review the knowledge about given genes using Indigo and its central concept of 'neighbourhood'.

3.1. Proximity in the chromosome

In a first example, partially developed elsewhere [23] and to which complementary information is given here, we start from the neighbourhood: 'proximity in the chromosome' and choose for our study an operon, *cmk rpsA*, that is conserved in *E. coli* and *B. subtilis*. We use neighbourhoods available in Indigo, using proximity between genes according to the way they use codons, to further explore our knowledge about these genes and gene products.

Analysis of the *cmk* gene, showing that it codes for cytidylate kinase, leads us to the study of pyrimidine biosynthesis. We first find, in the Swiss-Prot entry, that another name of the gene was *mssA*, for a suppressor of *smbA* (now *pyrH*, coding for uridylate kinase), which itself is a suppressor of *mukB*, a gene required for chromosome segregation. Looking in Indigo at the diagrams of biosynthetic pathways, we discover that there seems to be some difficulty in synthesizing de novo CDP, because CTP is formed directly from UTP, without going through CDP. This is a problem because it is known that synthesis of deoxyribonucleotides begins with from ribonucleoside *di*phosphates not triphosphates as start points... This leads us to look for RNA degradation processes. We now find the following list of neighbours: *bla*, *cat*, *pyrF*, *dicB*, *hflB*, *ftsH*, *mrsACF*, *lpp*, *nusA*, *ompA*, *pcnB*, *pnp*, *rna*, *rnb*, *rnc*, *rne*/*ams*, *rph*, *trxA* and *metY*. Several of these genes can be eliminated before further study, because they obviously correspond to the specific case of their own mRNA degradation (*bla*, *cat*, *dicB*, *lpp*, *ompA*). We note, however, the presence of a gene of pyrimidine metabolism, orotidine monophosphate decarboxylase (*pyrF*), that we must add to the information that *cmk* and *pyrH* are related via *mukB*. This places three genes of pyrimidine metabolism together. We then look for the functions of the other genes of the neighbourhood and find out processes linked to RNA turnover: RNaseIII (*rnc*, *nusA*, *metY*) [24], RNase H (*rnh*), housekeeping RNases (*rna*, *rnb*), *pnp* (for polynucleotide phosphorylase), *pcnB* (polyA polymerase) or *rne* (for RNase E). For the latter, we find a neighbour: *cafA* (cytoplasmic axial filament protein gene). It has the following noteworthy function (Swiss-Prot entry): "could be involved in chromosome segregation and cell division. In the chromosome the *cafA* gene is next to genes controlling in an unknown way the shape of the cell, *mreBCD*. It may be one of the components of the cytoplasmic axial filaments bundles or merely regulate the formation of this structure. It has similarity to the N-terminal half of *E. coli* ribonuclease E". This is interesting, and may be related to chromosome segregation mediated by MukB. Now, looking for the PNPase protein, we find that it degrades mRNA from their 3′ end into NDPs and *not* NMPs as other RNases do. In addition, we find that it comprises an 'S1-box' characterised from the S1 ribosomal protein of *E. coli* (which contains four S1-boxes in its sequence). Putting all this information together, it cannot escape our attention that the name of the S1 gene is *rpsA*, the gene downstream from the *cmk* gene in the operon we are studying. Although this might have occurred just by chance (but this is unlikely because these genes are found together in very distant organisms, *B. subtilis* and *E. coli*) we can go on by studying together PNPase, cytidylate kinase and S1.

To make a long story short (summarised in [23]), the combination of neighbourhoods leads us to make several links that must be meaningful for the cell: a complex degrading mRNA, the 'degradosome' comprising PNPase, RNase E, enolase and polyphosphate kinase, producing NDPs; another putative complex comprising S1, and other gene products (polyA polymerase?), and RNases, producing NMPs; and, finally, a complex helping to allow chromosome segregation, comprising *mukB*, *cafA*, and perhaps linked, via RNAse E, to the degradosome... This also permits us to suggest that CDP synthesis is central to DNA metabolism, and that its biosynthesis should be reconsidered and studied as a priority, not only in bacteria, but also in eucaryotes. The situation is much less clear in *B. subtilis*, probably because most genes found in the various neighbourhoods we explore are 'y' genes (unknown genes) so that it becomes very difficult to extract appropriate correlations. The *rpsA*-like gene, *ypfD*, has a codon usage similar to that of genes involved in nucleotide biosynthesis *ndk* (nucleoside diphosphokinase), and *guaB* (inosine dehydrogenase) but also *clpX* (coding for a chaperon-like protein), *gyrA* (coding for DNA gyrase), and two membrane proteins (*atpE* and *qoxC*). Because it has been experimentally demonstrated that there is no ribosomal S1 protein in *B. subtilis* [25,26], the function of the product of *ypfD* must differ from the recognised function in *E. coli*. It could, therefore, be part of a complex (that could be linked to the membrane) that processes mRNAs. In particular, this may be an example of the scaffolding of RNA molecules, using the counterparts of chaperonins, i.e. RNA chaperons. Several other proteins in *B. subtilis* possess S1-boxes (including the *pnpA* product): their genes are *yabR*, *ydcJ*, *yugI*, *yvaJ*. Clearly, as the function of more genes will be known, the story that can be built up using neighbourhoods will improve dramatically (Fig. 4).

3.2. Proximity in a cell compartment

In a second example, we explored the origin and

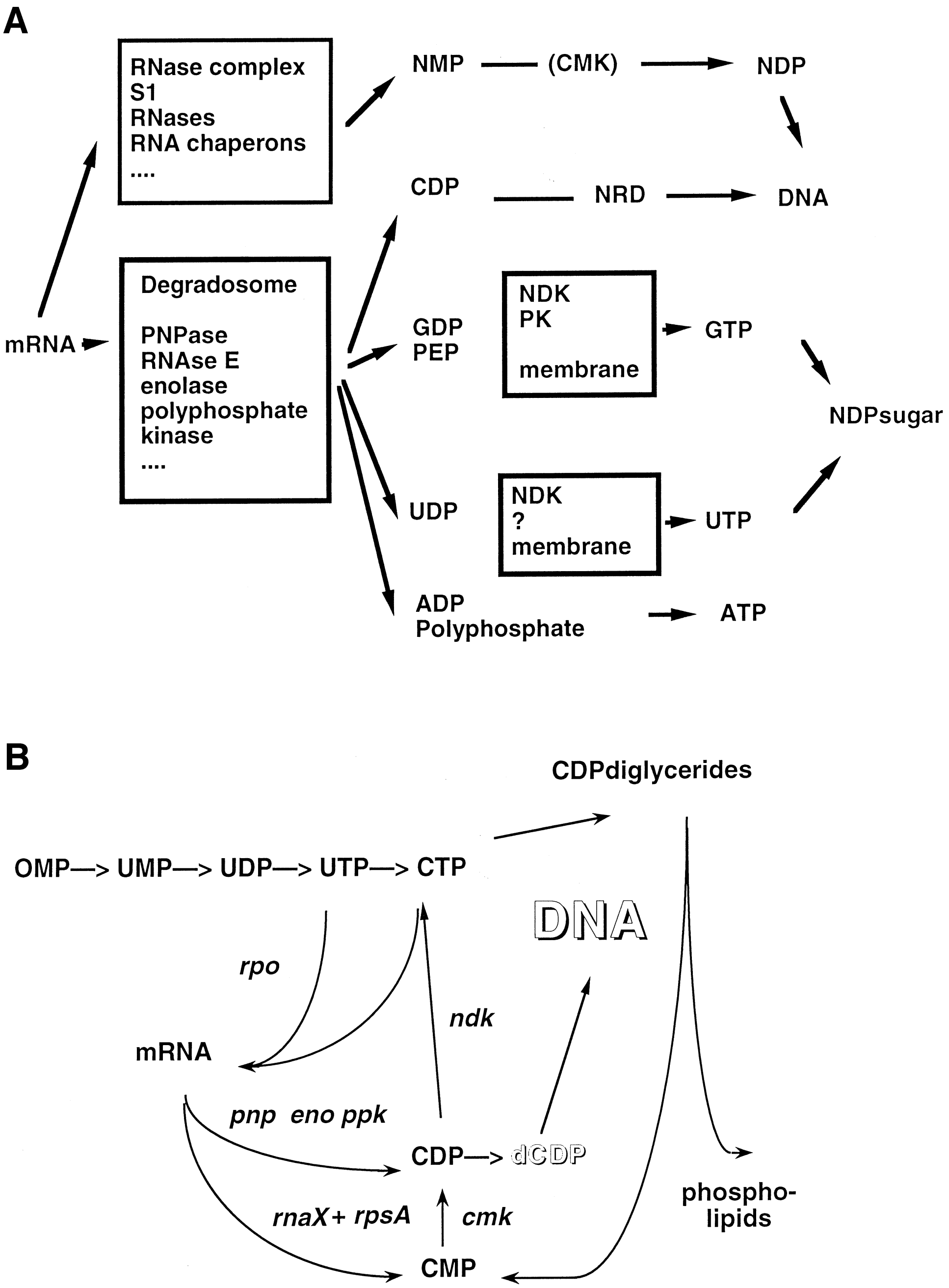


Fig. 4. The *cmk rpsA* CDP synthesising operon. Note that this supposes the existence of multienzymatic protein complexes, and the likely presence of RNA chaperons. (A) The pyrimidine biosynthetic pathways. (B) The involvement of mRNA turnover in nucleotide biosynthesis.

biosynthesis of a cell compartment, the outer membrane of *E. coli*. Because they comprise at least five compartments, Gram-negative bacteria pose a difficult challenge to protein targeting. Indeed, proteins must be synthesised in the cytoplasm, but although the majority are located in this compartment, some must also get into the inner membrane, into the periplasmic space, into the outer membrane or, finally for a few of them, be secreted in the external medium. Different mechanisms permit appropriate targeting. In particular, in a preferred pathway, a secretion machinery uses a signal peptide as a zip code to address protein to membrane or periplasmic compartments. However, although this seems relatively easy to consider for internalisation in the inner membrane or export to the periplasm, addressing to the outer membrane seems to be a significantly more difficult process to implement, because the proteins have to get through several compartments before they reach their target.

We made use of factorial correspondence analysis (FCA), a method which makes use of χ^2 distances for classification of objects without a priori knowledge of the classes [21], for analysis of the codon usage of genes directing synthesis of the major components of the outer membrane. Our goal was to analyse the molecules implicated in the cell compartments recognition. We chose the outer membrane as a test compartment because it is composed of a few types of abundant molecules: phospholipids, lipopolysaccharides (LPS), lipoproteins and porins. It was therefore easy, using the Indigo server to analyse the codon usage of the members of this compartment. Whereas porins all belong to a class having a highly biased codon usage (the same as that comprising the vast majority of ribosomal proteins), the LPS biosynthetic enzymes genes, *rfa* genes, display a completely different type of codon usage.

Combining this observation with the other neighbourhoods available in Indigo allowed us to propose that the outer membrane originates from different genomes, and has components synthesised at different locations inside the cell [27]. This work required interplay between codon usage analysis, knowledge about intermediary metabolism (in particular concerning LPS biosynthesis) and search for proteins homologous to porins. This allowed us to suggest the presence of a frameshift error in the sequence of a putative porin gene, and discuss about genes that have been involved in lateral transfer [20,28,29] and more generally about the pathways to speciation of Gram-negative bacteria [30,31].

3.3. Proximity in biochemical function

Let us now turn to other examples of neighbourhoods for which we do not have yet mechanistical explanations. Maintaining a codon usage bias supposes the existence of a strong and regular selection pressure. It is therefore interesting to investigate how genes are clustered according to this constraint (despite the fact that its origin is not yet understood). In particular, it should be rewarding to study how the components of the translation machinery are clustered. Of course it is impossible, simply by fixing a threshold distance between genes of differing codon

Table 1
Clusters of tRNA synthetase genes of *E. coli* and *B. subtilis* according to their codon usage

E. coli	*B. subtilis*
alaS, gltX, pheS, pheT, argS, serS, tyrS	*alaS, aspS, pheS pheT, glyQ, glyS, ileS, leuS, asnS, serS, valS, ytpR*
glyS, glyQ, ileS, metG, glnS	*cysS*
valS, proS, lysS, leuS, aspS	*gltX*
hisS	*hisS*
lysU	*hisZ*
cysS, trpS	*lysS*
	metS
	proS
	argS
	serS
	thrS
	thrZ
	trpS
	tyrS
	tyrZ

usage, to know whether this is statistically significant: this would require the comparison between the real genome and a model genome incorporating some of the constraints (base composition, dinucleotide composition etc). We therefore chose arbitrarily to consider the 10 or 20 neighbour genes, in each class of genes identified by cluster analysis [20]. The relevance in the corresponding formation of clusters will be justified if, and only if, this leads to a significant biological common property.

Most ribosomal protein genes have a highly biased codon usage, and indeed they constitute the core of class II genes, those genes that are expressed at a high level during exponential growth conditions [20]. But what about the genes for tRNA synthetases? In *E. coli*, 23 genes encode tRNA synthetases or subunits of tRNA synthetases. They can be selected using the 'functional category' neighbourhood. Going then to the codon usage neighbourhood for each of these genes, one creates a list of 20 neighbours, one for each gene. If now one collects overlapping clusters (i.e. clusters that have in common two or more genes), making 'cliques', we find in *E. coli* 6 cliques and in *B. subtilis* 15 cliques (Table 1).

A first observation can be made: in *E. coli* tRNA synthetases made of two different subunits (PheS PheT and GlyQ GlyS) have genes that are both members of an operon, and the codon usage in the operon is conserved. In general, the neighbours of the tRNA synthetase genes code for known functions in a high proportion (usually more than 75% of the genes are known). It is therefore remarkable that the proportion of unknown genes ('*y*' genes, and '*orf*' genes comprise more than 50% of its neighbours) is very high for *lysU* (which is a second lysine tRNA synthetase gene, linked to stress response [32]). One may also remark that *cysS* and *trpS* cluster together, as do their codons in the table of the genetic code.

There is no correlation in this clustering with the two families (I and II) of tRNA synthetases [33–35], as members of the two families are distributed into the six classes. No correlation either is found with the length or the quaternary structure (dimer, tetramer, etc.) of these enzymes. A weak correlation could be found either with metabolism (isoleucine, leucine and valine are in the same clique, as are phenylalanine and tyrosine) or with phylogeny (aspartate and lysine [36]).

The clustering does not seem to be correlated with what is currently known about tRNA synthetases. This asks an interesting question about the origin of the codon usage bias for these genes. One may therefore wonder whether the selection pressure acting on the codon usage is not due to the formation of macromolecular complexes comprising synthetases that are members of a clique and other proteins of apparently unrelated function. Evidence for compartmentalisation of certain components of the protein synthesis apparatus have been described in eucaryotic cells, in particular in mammalian cells [37], but no equivalent situation has been described in bacteria. However, it seems necessary that synthetases are present together in the vicinity of each ribosome. Apart from threonine tRNA synthetase, which poses a challenging problem (this may be related to the unusual and very complicated translation control of the synthesis of this enzyme [38,39]), it seems remarkable that synthetases corresponding to the rarest amino acids in proteins (cysteine, histidine and tryptophan) are not clustered with the cliques comprising the majority of synthetases. It should also be stressed that these amino acids are precisely those that are the most prompt to oxidation. It might be interesting for the cell to compartmentalise the synthetases in such a way that it has minimum risks to be charged by an oxidised amino

Table 2
Orthologous genes between *E. coli* and *B. subtilis*

Operons in *E. coli*	Corresponding independent genes in *B. subtilis* and location on the chromosome (kb)
araABD	*araA* (2948), *araB* (2946), *araD* (2944)
sdhABC	*sdhA* (2907), *sdhB* (2905), *sdhC* (2907)
glnHPQ	*glnH* (2802), *glnP* (2803), *glnQ* (2801)
ilvABCDN	*ilvA* (2293), *ilvB* (2896), *ilvC* (2893), *ilvD* (2301), *ilvN* (2894)
mreBCD	*mreB* (2861), *mreC* (2860), *mreD* (2859)
galETK	*galE* (3989), *galK* (3920), *galT* (3919)
thrABC	*hom* (3315), *thrB* (3312), *thrC* (3313)
cheBRY	*cheB* (1711), *cheR* (2380), *cheY* (1703)
murBCDEFG, mraY	*murB* (1592), *murC* (3048), *murD* (1588), *murE* (1585), *murF* (508), *murG* (1590), *mraY* (1587)

The genes displayed in the table are clustered in an operon in *E. coli* and independent in *B. subtilis*.

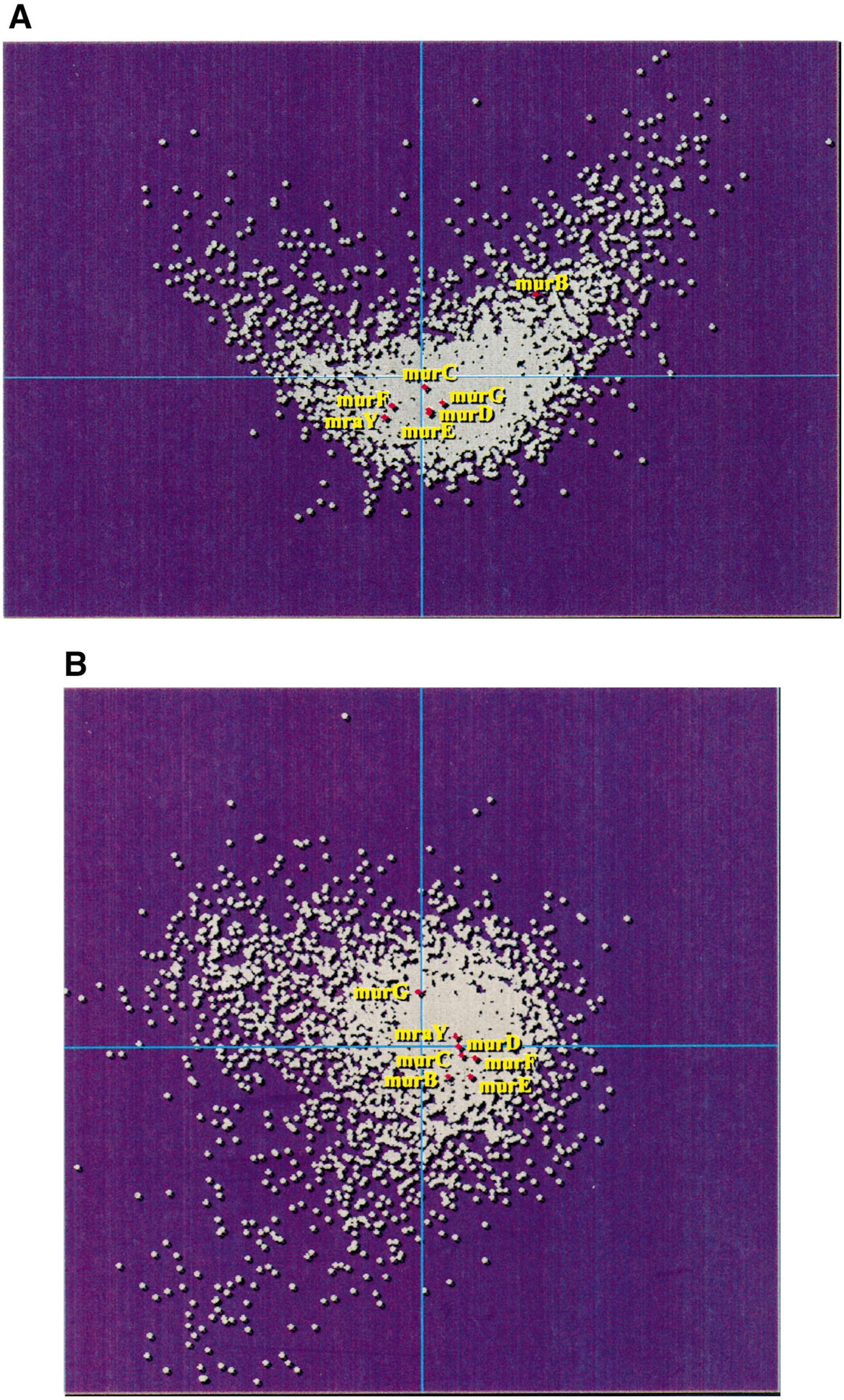


Fig. 5. Factorial correspondence analysis of the codon usage for seven genes (*murBCDEFG mraY*) in *E. coli* and *B. subtilis*.

Table 3
A family of paralogous genes in *E. coli* (similarity *Z*-score = 14)

Gene	Product length (aa)	Function	Induction	Product localisation	Similarity reported in Swiss-Prot	Chromosome localisation
acs	652	Acetyl CoA synthase	Inducible glycerol without glucose		Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	4285
entE	536	2,3-Dihydroxybenzoate AMP- ligase	Constitutive	Inner membrane	Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	625.3
entF	1293	Enterocheline synthase F	Constitutive	Inner membrane (potential)	Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	613.4
menE	451	OSB-CoA-ligase	Constitutive		Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	2373
aas	719	Acyl ACP synthase	Constitutive	Inner membrane	Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	2974
fadD	561	Long chain fatty acid CoA ligase	Constitutive	Inner membrane associated (potential)	Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	1887.8
caiC	522	Carnithine CoA ligase	Constitutive	Inner membrane (potential)	Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	37.8
ydiD	566	Hypothetical protein	?		Similarity to others enzymes which act via an ATP-dependent covalent binding of AMP to their substrate	1781
*orf*0086	628	Unknown	Inducible?		40% Similarity with ACSA_ALCEU	1884

The corresponding FCA of their codon usage is displayed in Fig. 6.

acid. This may suggest different compartmentalisation of these enzymes with respect to ribosomes.

The same study performed with *B. subtilis* genes (25 genes) gives very different results, since we find one large clique (*alaS asnS aspS ileS leuS pheS pheT serS valS ytpR*), using a 5% cutoff for the maximum distance of a neighbour to each gene analysed. One could perhaps add the *glyQ glyS* couple to this clique, because these genes are located into neighbourhoods comprising, as we did in *E. coli*, the 20 genes proximal to at least one member of the clique. In contrast, all other tRNA synthetase genes remain

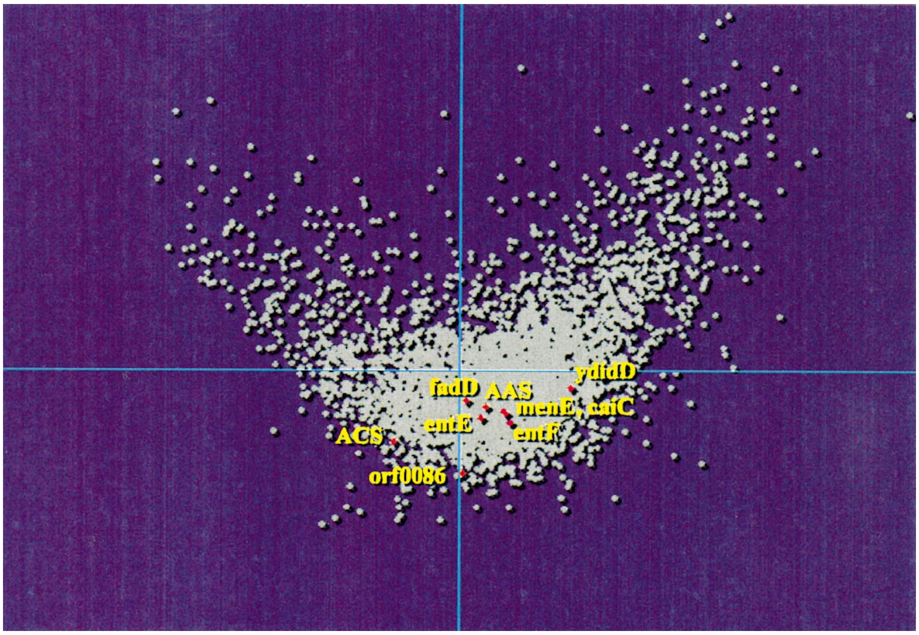


Fig. 6. The paralogues of ligases acting via an ATP-dependent covalent binding of AMP to their substrate are clustered in the FCA cloud of points of the gene codon usage in *E. coli*. *acs* and *orf0086* exceptions are discussed in the text.

isolated. The situation for *lysS*, *metS*, *proS* is difficult to account for: they are not very far from the clique comprising the bulk of the synthetases, and they may indeed belong to it (as stated above the 5% distance threshold is arbitrarily chosen because we cannot identify a significant statistical threshold in the absence of a theoretical model of a bacterial chromosome [40]). It must be remarked here that the organisation of tRNA synthetases in metabolism is different in *E. coli* and *B. subtilis*. For example, three synthetases exist as doublets (*hisSZ*, *thrSZ* and *tyrSZ*) and the reason for this is not known (Diaz-Lazcoz and co-workers interpret this as a sign of early gene duplication ([41]). Glutamine tRNA synthetase does not exist in *B. subtilis*, and in fact glutamyl tRNA synthetase (GltX) charges both tRNA^{glu} and tRNA^{gln} with glutamate. The second is subsequently amidated by homeotopy ([42]). As in *E. coli* Cys, Trp and His are rare amino acids in proteins. The most noticeable exception is ArgS. In contrast to all the other tRNA synthetase genes, that are not very distant from each other in the codon usage plot, the *argS* gene is located in a codon usage region very far from the bulk, suggesting a special regulation (and perhaps distribution of the enzyme in the cell). Its gene is located downstream of gene *ywiB* (with no known function) within an operon.

The regulation of tRNA synthetase synthesis is very different in *E. coli* and in *B. subtilis*. In the latter, there exists for most synthetases (including the *tyrS tyrZ* and *thrS thrZ* genes), a regulation at the transcription level by binding of the cognate charged tRNA to a mRNA leader sequence (T-box tRNA synthetase genes [43]). *gltX* and *cysS* are part of a common operon [44], but the T-box mediated regulation is between *gltX* and *cysS*, so that we cannot put this regulatory feature forward as a common point to the tRNA synthetase clustering in *B. sub-*

Table 4
Genes involved in the aromatic amino acids biosynthesis pathway in *E. coli* and *B. subtilis*

Pathway	*E. coli*	*B. subtilis*
General	*aroABCDEFGHKL*	*aroABCDEFHI*
Phenylalanine	*pheA, tyrB*	*pheA, pheB*
Tyrosine	*tyrA, tyrB*	*tyrA*
Tryptophan	*trpABCDE*	*trpABCDEF*

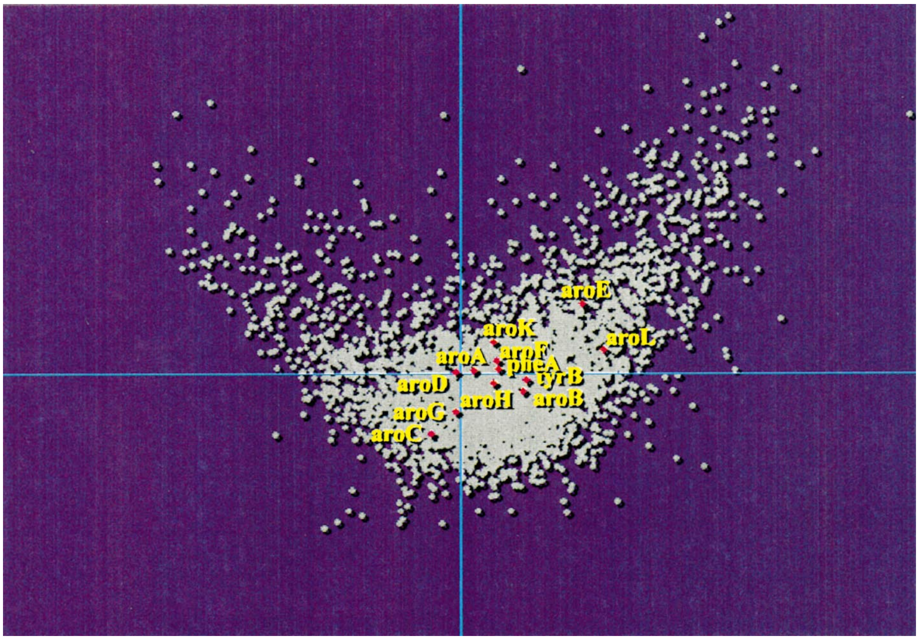


Fig. 7. Distribution of genes involved in phenylalanine biosynthesis in the codon usage FCA cloud of points.

tilis. The *argS* gene expression does not seem to be regulated by this common process. It is therefore worth considering that codon usage may reflect a correlation between geographical location on the chromosome and the architecture of the cell.

The tRNA synthetase genes are distributed differently in the chromosome of *E. coli* and *B. subtilis*. Genes that are members of the same clique are not close to each other in the chromosome. This led us to test for the periodicity of the distances computed between genes of the same clique. As a preliminary result, it appeared that tRNA synthetase genes are not randomly distributed (data not shown). We must, however, be very cautious at this point because statistics on finite sets are notoriously difficult to perform, and that the set of tRNA synthetase genes is very small in any case.

Finally, it seems interesting to correlate our codon usage classification with phylogenic considerations. Is a trace of the evolution of genes and species marked in the codon usage of tRNA synthetases? Considering codon utilisation, our observations are consistent with those of Diaz-Lazcoz et al. [41]. Using a similar approach, these authors showed that the genes encoding a tRNA synthetase might reflect the outcome of evolution both by gene duplication and by lateral transfer. In *E. coli* and *B. subtilis*, they showed that most genes cluster in a single codon usage category (using distances between genes of similar codon bias somewhat larger than the one we used here). When they found exceptions, they interpreted their observation as reflecting an important influence of horizontal transfer. We should stress, however, that, although lateral transfer is certainly important in evolution (and may, for a significant period of time, result in a codon bias that differs from that of the bulk of the biosynthetic system of an organism [20]), it seems unlikely, in view of the importance of tRNA synthetases in translation (except in the case of gene duplication), that a tRNA synthetase gene could have been incorporated in the genome recently enough to keep its original bias, unless it is associated to a function that is regularly prompted. In contrast, there exists another selection pressure that may play a fundamental role in maintaining a codon bias: mRNA molecules could be translated in specific regions of the cell, where the average codon usage might differ from the core [27]. In particular, this would be the case if a gene product (here, a tRNA synthetase) would be part of a complex with other enzymes. In this respect, it will be very important to further analyse the situation of *argS* in *B. subtilis*.

3.4. *Proximity in evolution: orthologous genes and operon conservation between E. coli and B. subtilis*

Using the information present in Indigo, we extracted an exhaustive list of genes orthologous between *E. coli* and *B. subtilis*. In this list (1008 genes) we found 328 genes that are conserved as an operon or a fragment of operon in both genomes. Thirty-three genes included in an operon in *E. coli* are distributed and transcribed as independent entities in *B. subtilis*. The converse is true for 24 genes that are clustered in an operon in *B. subtilis* and independent in *E. coli*. The 618 remaining genes should give independent transcripts in both cases. Let us consider the genes that are clustered as an operon in *E. coli* and not in *B. subtilis* (Table 2). When involved in related functions, their codon usage is the same within the considered organism (but of course differs between *E. coli* and *B. subtilis*, which do not use the genetic code in the same way), whether or not the genes are part of an operon. As an example, the codon usage for a set of seven genes (*murBCDEFG mraY*) is illustrated in Fig. 5. One observes in the figure that *E. coli* and *B. subtilis* behave in a quite similar way, in that in each case, genes cluster together: the only exceptions are *murB* in *E. coli* and *murG* in *B. subtilis*. This led us to measure the distance in the chromosome between the corresponding genes in *B. subtilis* in order to see, when genes belong to independent transcription units, whether they displayed a particular distribution. Our preliminary results suggest that these genes (compared to a randomly distributed gene sample) are not distributed randomly in the chromosome. At the time of this review, we cannot completely account for these results, but this can be taken as a support of the hypothesis that the map of the cell is correlated to the map of the chromosome [45].

A second example of proximity in evolution shows how, using neighbourhood exploration, the review capability of Indigo allows one to predict gene functions and/or expression conditions. In this example, we explored the significance of a family of homologous proteins (paralogues: this represents a very familiar relationship between sequences analysed by scientists when they have access to genome sequences). The main idea here is to look and use the codon usage FCA to see whether the genes comprised in a given functional family identified by paralogy is homogeneous or not, with respect to synonym codon utilisation. The chosen class was the family in *E. coli* of ligases which act via an ATP-dependent covalent binding of AMP to their substrate. The family is composed of paralogous genes showing a high level of sequence similarity. Members of the family were selected using neighbourhood homology in Indigo, in such a way that each gene in the family was similar with at least one of the others (connex family, or clique). Computing a *Z*-score between each pair of amino acid sequences of potential paralogous genes, using the Smith and Waterman algorithm, and a base line formed with 100 sequences generated from one of the pair, we retained pairs that were characterised by a *Z*-score higher than 14, ensuring a high level of similarity. With this score, we built a family comprising 9 genes (Table 3).

Fig. 6 shows that whatever the length of the corresponding product (451–1293 amino acid residues), the localisation of the genes in the chromosome map, or the metabolic pathway in which the gene is implicated, the codon usage of the homologous genes clustered in the family in study is similar, with two exceptions. The first one, *orf0086* is annotated as an 'open reading frame' (not a coding sequence). The start site of the corresponding coding sequence is not clearly defined, and the gene has not been formally identified as a bona fide gene. Its sequence uses codons differently as compared to the other sequences in the dataset. This can be used as an element to discuss its potential function. Sequence homology suggests that ORF0086 is an ATP-dependent ligase of the same type as that of the others members of the family. At this stage of the investigation, we cannot say much more (see below). The second exception is gene *acs*, that codes for an acetylCoA synthase. Why would this gene use synonym codons differently from the other members of the family, and what would be the consequences of a different codon bias on its function and/or expression conditions? Using the literature data linked in Indigo to this gene we propose the beginning of an explanation. Clark and Cronan have reviewed the literature on metabolism of acetate in *E. coli*. In this organism *ackA*$^-$ and *pta*$^-$ mutants can incorporate acetate when grown on glycerol, but fail to do so when grown in a medium supplemented with glucose as the carbon source [46]. They have concluded that, in addition to the major pathway to transform acetate into acetyl-phosphate *E. coli* contains a second pathway. The latter pathway is not constitutive, but submitted to catabolite repression. It involves the inducible acetylCoA synthase encoded by the *acs* gene. At this point we note that, when known, all genes in the family we are studying have a *constitutive* phenotype (Table 4). We propose, therefore, that the reason underlying the exception in the *acs* codon bias could be explained by its particular expression conditions: its expression is regulated, not constitutive. A further argument substantiates this hypothesis. We have just seen that *orf0086* also has a particular codon usage. Its product, ORF0086, shows 40% identity with ACSA_ALCEU. This protein of *Alcaligenes eutrophus* is an acetylCoA synthase: its gene *acsA* is the orthologue of *acs* in *E. coli*. It therefore seems plausible that *acsA* is also inducible by glycerol in this organism. In turn, we might accept as a working hypothesis that the expression of *orf0086* is also inducible. In the framework of an *E. coli* genome functional exploration programme (such as the one for yeast, Eurofan [47]) this would permit us to propose that ORF0086 is an ATP-dependent ligase and that its expression is not constitutive, and should be studied in the appropriate experimental conditions. This example illustrates how neighbourhood analysis via Indigo allows us to substantiate or predict gene expression conditions.

3.5. Proximity in metabolic pathways

As another of its capabilities, Indigo allows one to select genes having products involved in the same metabolic pathway. As an example, we studied aromatic amino acids biosynthesis. The corresponding genes are listed in Table 4. As stated above, we expect that the expression level of a gene (reflecting that of the parallel metabolic function) exerts a selection pressure on the corresponding codon biases. Put differently: is the codon usage of a gene correlated to the biological function of its product? As we can see in Fig. 7, the distribution of genes along a given metabolic pathway (the phenylalanine biosynthesis pathway) in the cloud of points does not appear to be random. Remarkably, the genes of the pathway seem to be distributed along a straight line. Because the plot is the projection of a 61-dimensional space (the codon space), this was verified: indeed the points do not deviate much from a straight line. Although not yet completely understood in structural and molecular terms, this observation provides a major argument in favour of a correlation between synonymous codon utilisation in a gene and biological function of its product. This observation was further extended and substantiated when studying other amino acids biosynthetic pathways (data not shown). The correlation remains true when genes are clustered into an operon (this is the case, for example, of histidine biosynthesis), or distributed along the chromosome (aromatic amino acids biosynthesis). We have found no evidence for a correlation with the length of genes or substrates-products order in the pathway that could explain this particular codon usage.

4. Conclusion

The explosion of genomics, that is even more important than the explosion of biological literature, asks for new means to make the associated knowledge available to most scientists in a fruitful way. This is a compelling incentive to rethink the way to collect, manage and present experimental data related to genomics, together with their associated in silico analyses and observations. As a consequence, we are facing a unique opportunity to realise that the usual hypothetico-deductive approach to biological questions, although efficient to refine pre-existing knowledge, is utterly insufficient to permit discovery. We propose that organisation by neighbourhood, of in vivo, in vitro and in silico data into a general structure allowing the scientist to combine different points of view on the facts, objects and relationships he/she is interested in, is an efficient way to generate new knowledge, and hence to lead to discovery. Annotating genomes should benefit immensely from such an approach if it can be constructed in an efficient and user-friendly way. We hope that the Indigo server (http://indigo.genetique.uvsq.fr) will help progress in this direction.

Acknowledgments

This work was supported by the *Bacillus subtilis* Industrial Platform (BACIP), Groupement d'Etudes et de Recherches sur les Génomes (GREG), Université de Versailles-Saint-Quentin and Institut Pasteur.

References

[1] Benson, D., Boguski, M., Lipman, D., Ostell, J. and Ouellette, B. (1998) GenBank. Nucleic Acids Res. 26, 1–7.

[2] Stoesser, G., Sterk, P., Tuli, M., Stoehr, P. and Cameron, G. (1997) The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 25, 7–14.

[3] Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T. (1998) DNA Data Bank of Japan at work on genome sequence data. Nucleic Acids Res. 26, 16–20.

[4] Bairoch, A. and Apweiler, R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res. 26, 38–42.

[5] Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. (1998) SGD: *Saccharomyces* Genome Database. Nucleic Acids Res. 26, 73–79.

[6] Gaasterland, T. and Sensen, C. (1996) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie 78, 302–310.

[7] Kröger, M. and Wahl, R. (1998) Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC. Nucleic Acids Res. 26, 46–49.

[8] Kunisawa, T., Nakamura, M., Watanabe, H., Otsuka, J., Tsugita, A., Yeh, L., George, D. and Barker, W. (1990) *Escherichia coli* K12 genomic database. Protein Seq. Data Anal. 3, 157–162.

[9] Médigue, C., Viari, A., Hénaut, A. and Danchin, A. (1993) Colibri: a functional data base for the *Escherichia coli* genome. Microbiol. Rev. 57, 623–654.

[10] Slonimski, P. and Brouillet, S. (1993) A data-base of chromosome III of *Saccharomyces cerevisiae*. Yeast 9, 941–1029.

[11] Moszer, I., Glaser, P. and Danchin, A. (1995) SubtiList: a relational database for the *Bacillus subtilis* genome. Microbiology 141, 261–268.

[12] Benson, D., Boguski, M., Lipman, D. and Ostell, J. (1994) GenBank. Nucleic Acids Res. 22, 3441–3444.

[13] Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. and Krummenacker, M. (1998) EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res. 26, 50–53.

[14] Overbeek, R., Larsen, N., Smith, W., Maltsev, N. and Selkov, E. (1997) Representation of function: the next step. Gene 191, GC1–GC9.

[15] Selkov, E.J., Grechkin, Y., Mikhailova, N. and Selkov, E. (1998) MPW: the Metabolic Pathways Database. Nucleic Acids Res. 26, 43–45.

[16] Bono, H., Ogata, H., Goto, S. and Kanehisa, M. (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res. 8, 203–210.

[17] Neidhardt, F.C. (1996) *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology, ASM Press, Washington, DC.

[18] Médigue, C., Gascuel, O., Soldano, H., Hénaut, A. and Danchin, A. (1991) From data banks to data bases. Res. Microbiol. 142, 913–916.

[19] Myhill, J. (1952) Some philosophical implications of mathematical logic I. Three classes of ideas. Rev. Metaphys. 6, 165–198.

[20] Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. 222, 851–856.

[21] Hill, M.O. (1974) Correspondence analysis: a neglected multivariate method. Appl. Stat. 23, 340–353.

[22] Sneath, P. (1957) The application of computers to taxonomy. J. Gen. Microbiol. 17, 201–226.

[23] Danchin, A. (1997) Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. DNA Res. 44, 9–18.

[24] Régnier, P. and Grunberg-Manago, M. (1990) RNase III cleavages in non-coding leaders of *Escherichia coli* transcripts control mRNA stability and genetic expression. Biochimie 72, 825–834.

[25] Farwell, M., Roberts, M. and Rabinowitz, J. (1992) The effect of ribosomal protein S1 from *Escherichia coli* and *Micrococcus luteus* on protein synthesis in vitro by E. coli and Bacillus subtilis. Mol. Microbiol. 6, 3375–83.

[26] Isono, K. and Isono, S. (1976) Lack of ribosomal protein S1 in *Bacillus stearothermophilus*. Proc. Natl Acad. Sci. USA 73, 767–770.

[27] Guerdoux-Jamet, P., Hénaut, A., Nitschké, P., Risler, J. L. and Danchin, A. (1997) Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res. 4, 257–265.

[28] Boyd, E. and Hartl, D. (1998) Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J. Bacteriol. 180, 1159–1165.

[29] Lawrence, J. (1997) Selfish operons and speciation by gene transfer. Trends Microbiol. 5, 355–359.

[30] Taddei, F., Vulic, M., Radman, M. and Matic, I. (1997) Genetic variability and adaptation to stress. EXS 83, 271–290.

[31] Moreno, E. (1997) In search of a bacterial species definition. Rev. Biol. Trop. 45, 753–771.

[32] Saluta, M. and Hirshfield, I. (1995) The occurrence of duplicate lysyl-tRNA synthetase gene homologs in *Escherichia coli* and other procaryotes. J. Bacteriol. 177, 1872–1878.

[33] Cusack, S. (1997) Aminoacyl-tRNA synthetases. Curr. Opin. Struct. Biol. 7, 881–889.

[34] Delarue, M. (1995) Aminoacyl-tRNA synthetases. Curr. Opin. Struct. Biol. 5, 48–55.

[35] Landes, C., Perona, J., Brunie, S., Rould, M., Zelwer, C., Steitz, T. and Risler, J. (1995) A structure-based multiple sequence alignment of all class I aminoacyl-tRNA synthetases. Biochimie 77, 194–203.

[36] Gatti, D. and Tzagoloff, A. (1991) Structure and evolution of a group of related aminoacyl-tRNA synthetases. J. Mol. Biol. 5, 557–568.

[37] Cerini, C., Kerjan, P., Astier, M., Gratecos, D., Mirande, M. and Semeriva, M. (1991) A component of the multisynthetase complex is a multifunctional aminoacyl-tRNA synthetase. EMBO J. 10, 4267–4277.

[38] Romby, P., Caillet, J., Ebel, C., Sacerdot, C., Graffe, M., Eyermann, F., Brunel, C., Moine, H., Ehresmann, C., Ehresmann, B. and Springer, M. (1996) The expression of E. coli threonyl-tRNA synthetase is regulated at the translational level by symmetrical operator–repressor interactions. EMBO J. 15, 5976–5987.

[39] Comer, M., Dondon, J., Graffe, M., Yarchuk, O. and Springer, M. (1996) Growth rate-dependent control, feedback regulation and steady-state mRNA levels of the threonyl-tRNA synthetase gene of *Escherichia coli*. J. Mol. Biol. 261, 108–124.

[40] Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I. and Danchin, A. (1996) Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. J. Mol. Biol. 257, 574–585.

[41] Diaz-Lazcoz, Y., Aude, J.-C., Nitschké, P., Chiapello, H., Landès-Devauchelle, C. and Risler, J.-L. (1998) Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases, submitted for publication.

[42] Danchin, A. (1990) Homeotopic transformation and the origin of translation. Prog. Biophys. Mol. Biol. 54, 81–86.

[43] Rollins, S., Grundy, F. and Henkin, T. (1997) Analysis of cis-acting sequence and structural elements required for antitermination of the *Bacillus subtilis tyrS* gene. Mol. Microbiol. 25, 411–421.

[44] Gagnon, Y., Breton, R., Putzer, H., Pelchat, M., Grunberg-Manago, M. and Lapointe, J. (1994) Clustering and co-transcription of the *Bacillus subtilis* genes encoding the aminoacyl-tRNA synthetases specific for glutamate and for cysteine and the first enzyme for cysteine biosynthesis. J. Biol. Chem. 269, 7473–7482.

[45] Danchin, A. and Hénaut, A. (1997) The map of the cell is in the chromosome. Curr. Opin. Gen. Dev. 7, 852–854.

[46] Clark, D. and Cronan Jr., J. (1996) in *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology, Vol. 1 (Neidhardt, F.C., Ed.), pp. 343–357. ASM Press, Washington, DC.

[47] Dujon, B. (1998) European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. Electrophoresis 19, 617–624.