

# Everything you always wanted to know about the prokaryotic world

## Cross-linking hundreds of genomes at the EBI

Twenty years ago, researchers decoded the first genomes of very small organisms – viruses, followed by the much larger genomes of bacteria, then finally yeast, flies, humans, and other animals. In the meantime, scientists have the complete DNA sequences of over 800 viruses, 140 bacteria, and 23 members of our own evolutionary family, the eukaryotes. The gene sequences have been stored in EMBL-Bank and other primary databases (GenBank in the US and the DDBJ in Japan), accompanied by information such as where genes are located in the genomes and what proteins they are known to encode.

Research accumulates new information about genes and proteins and the databases are revised to keep up, but the original genomic sequence is seldom updated. And researchers may still find it difficult to access everything that is known about a particular gene. “What’s been missing,” says Paul Kersey of the EBI, “is a comprehensive and standardized resource that links this information.”

Two years ago, the European Commission funded a major project called TEMBLOR, coordinated by Graham Cameron at the EBI; the major aim of the project is to create new tools that integrate biological data. Under the grant Paul has headed the development of a new resource called *Genome Reviews* that attempts to fill in some of the gaps.



“Genome Reviews enhances the sequences submitted to EMBL-Bank (or GenBank or the DDBJ) by cross-referencing them to many other resources,” Paul says. “For example, it’s now possible to hop straight from the genomic sequence to the proteins encoded there, maintained in UniProt, or information about other proteins that share a particular motif or domain; those are found in InterPro.”

Proteins often undergo modifications after they have been produced, and the resource gives researchers the ability to track them.

“In total, the number of cross-references has increased from roughly 650,000 in the corresponding part of EMBL-Bank to 2.5 million in Genome Reviews,” Paul says. “Perhaps most importantly, we’ve invested a great deal of effort in standardizing annotations. Original submissions were often very inconsistent in the way they described things. Standardization makes it much easier to compare data across genomes. We’ve also added 4.5 million ‘evidence’ tags showing where data came from – that’s a key to evaluating the quality of information.”

Under Rolf Apweiler, the EBI is developing a user interface called Integr8, giving researchers access to the information in Genome Reviews and the other EBI data resources. The standardization effort carried out by Paul and his team will make it much easier for scientists to understand that information in an integrated way. The resource complements other projects like Ensembl, which offers many of the same possibilities for the genomes of humans, mice, flies, and other eukaryotic animals.