Comparative Genomics

A Powerful New Tool in Biology

Anand K Bachhawat

Anand K Bachhawat is a scientist at the Institute of Microbial Technology, Chandigarh, and works in the area of yeast genetics, genome-wide analysis and molecular biology.

An important hallmark of biological research is the aspect of 'comparisons'. As the complete genome sequences of numerous organisms have become available, the emphasis in biology has shifted to comparisons at the genome level. Indeed, the last few years have witnessed an exponential rise in the number of organisms whose complete genome has been sequenced, and we are still climbing up the graph. The present article, a primer, explains how one can extract a great deal of information from such analyses that is of great value in our research. The subject of comparative genomics impinges on evolutionary biology and phylogenetic reconstructions of the tree of life, drug discovery programs, function predictions of hypothetical proteins and genes, regulatory motifs and other non-coding DNA motifs, and genome flux and dynamics. Finally the article describes how the information one can extract from a comparative analysis of genomes depends to a large extent, on the specific aspect of the genomes that is being compared and the phylogenetic distances of the organisms involved.

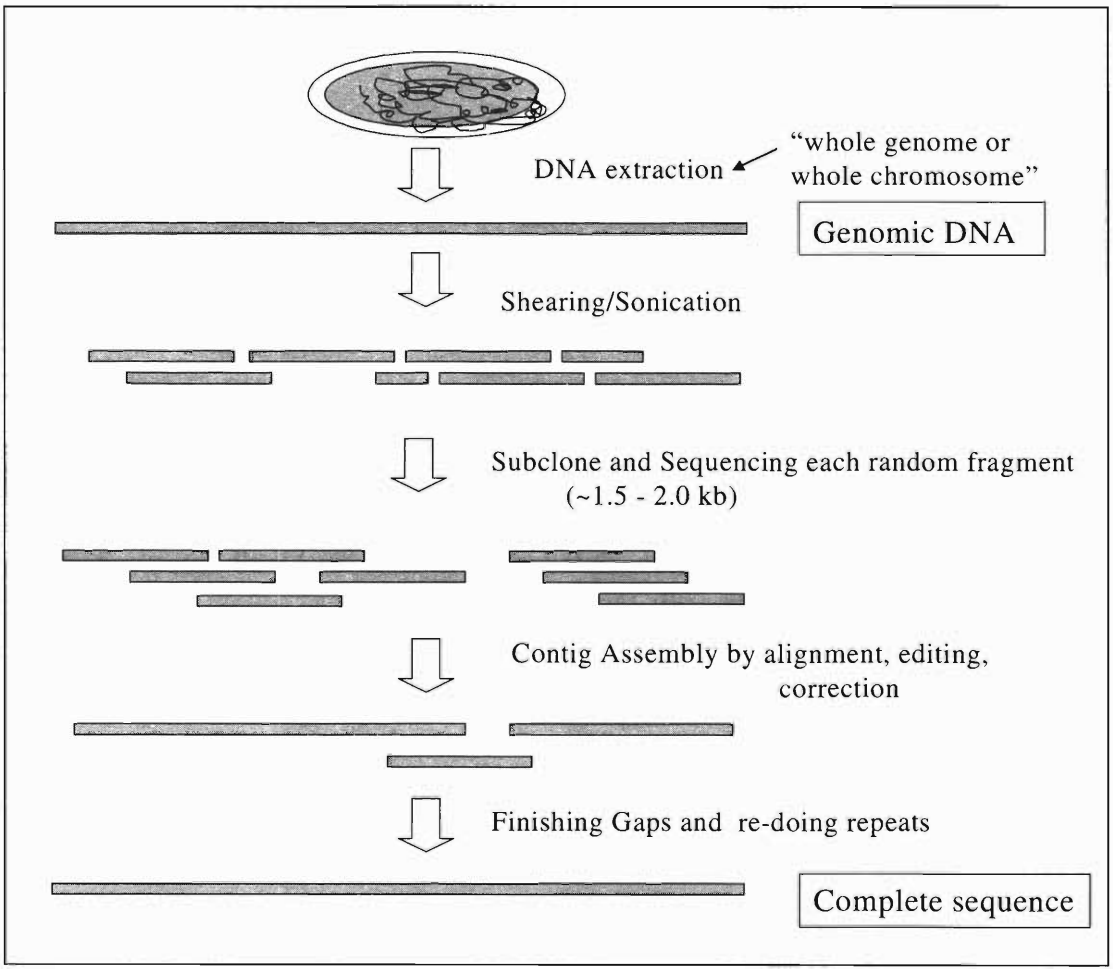
Introduction

Comparison is an important aspect of all biological research. The 'control experiment' or the 'control data' is often the all-important data in a biological experiment. In the early days of biological research, the comparisons were at the morphological and physiological level. These became more biochemical, and when the sequencing of proteins and DNA became possible, the comparisons reached a new level – they became molecular. Now as the complete genome sequences of several organisms has become available, the emphasis has shifted to comparisons at the whole genome level. Indeed, the last few years have witnessed an

Keywords
Comparative genomics, orthologue identification, molecular phylogeny, horizontal gene transfer, genome flux.

exponential rise in the number of organisms whose complete genome has been sequenced. The increased speed in whole genome sequencing in recent years is a result of the 'random shotgun strategy' (*Figure* 1) that is more amenable to automation, as well as improved instrumentation and better software and expertise in assembling sequences. Beginning with the complete genome sequence of the bacterial pathogen *Haemophilus influenzae* that was completed in 1995, several hundered complete genome sequences of different organisms are now available in the public domain (*Table* 1). With genome comparisons, have we reached a near ultimate level of comparison in biology? Or can comparisons go deeper? The answer perhaps lies in understanding what

Figure 1. The random whole genome (or chromosome) shotgun sequencing strategy was first introduced by Celera Genomics. In the earlier hierarchal strategy, larger clones (~150 kb) were first created in either bacterial artificial chromosomes (BACs) or Yeast artificial chromosomes (YACs), mapped carefully and then subjected to sequencing. (Kb=Kilobase)



Some genomes and their sizes. The smallest genome belongs to the obligate archaebacterial symbiont, *Nanoarchaeum equatans* which grows in association with *Ignococcus* spp. The smallest eukaryote is the eukaryotic parasite, *Encephalitozoon cuniculi*, and the smallest bacterial genome is that of the obligate intracellular parasite, *Mycoplasma genitalium*.

Genome sequencing projects (source: http://genomesonline.org) Eukaryote(E) sequences completed:41, Eukaryote sequences ongoing: 494. Eubacterial (B) sequences completed: 277, Eubacterial sequences ongoing: 933. Archaeal (A) sequences completed: 25, Archaeal sequences ongoing: 58

	Genes/ORFs	Genome Size
Nanoarchaeum equtans (A)	552	0.49 MB
Methanococcus jannaschii(A)	1682	1.7 MB
Methanosarcina acetivorans (A)	4540	5.8 MB
Mycoplasma genitalium(B)	468	0.58 MB
Helicobacter Pylori(B)	1590	1.7 MB
Escherichia coli(B)	4668	4.6 MB
Bacillus subtilis(B)	4221	4.2 MB
Mycobacterium tuberculosis (B)	3974	4.4 MB
Nostoc punctiforme (B)	7432	9.8 MB
Encephalitozoon Cuniculi (E)	1997	2.5 MB
Saccharomyces cerevisiae(E)	6500	12.1 MB
Caenorhabditis elegans(E)	19,000	97.1 MB
Drosophila melanogaster(E)	14,000	137 MB
Arabidopsis thaliana(E)	25,000	115 MB
Homo sapiens (E)	35,000	3000 MB

Table 1. Some genomes and their sizes.

comparison at the whole genome level is all about, and what it can really reveal.

Comparative genomics involves comparing any (or all) of a myriad aspects of the genomes of the organisms subjected to comparison. One can extract a great deal of information from such analyses that is of great value in evolutionary biology, genome dynamics and the phylogenetic reconstructions of the tree of life, drug discovery programs and function predictions of hypothetical proteins. In addition to this, identification of gene regulatory motifs, intronic regions and splice sites, as well as

other DNA motifs in non-coding regions are also set to benefit extensively from comparative analysis of genomes.

To be able to discuss the different aspects of comparative genomics, it is important to recapitulate the precise meaning of some of the terminologies that have evolved with the growth of the subject. Much confusion among researchers arises from a lack of clarity of the precise meaning and implication of these terms. Some important terms are described in *Box* 1.

Box 1.

Homology is the relationship of any two characters (such as two proteins that have similar sequences) that have descended, usually through divergence, from a common ancestral character. Two proteins can thus be either 'homologous' or 'non-homologous' and the term 'partial homology' is an incorrect usage. (In contrast, proteins can be 'partially similar' or 'partially identical' and can have a percentage similarity or identity between them.)

Homologues are thus components or characters (such as genes/proteins with similar sequences) that can be attributed to a common ancestor of the two organisms during evolution. Homologues can either be orthologues, paralogues, or xenologues (schematically depicted in *Figure* 2).

Orthologues are homologues that have evolved from a common ancestral gene by speciation. They usually have similar functions.

Paralogues are homologues that are related or produced by duplication within a genome. They often have evolved to perform different functions.

Xenologues are homologues that are related by an interspecies (horizontal transfer) of the genetic material for one of the homologues. The functions of the xenologues are quite often similar.

Analogues are non-homologous genes/proteins that have descended convergently from an unrelated ancestor (this is also referred to as 'Homoplasy'). They have similar functions although they are unrelated in either sequence or structure. This is a case of 'non-orthologous gene displacement'. In complete contrast to what most researchers instinctively believed, genome comparisons have revealed that this 'non-orthologous gene displacement' occurs at surprisingly high frequency.

Horizontal (Lateral) Gene Transfer is the movement of genetic material between species (or genus) other than by vertical descent. In bacteria this process occurs by either natural transformation, conjugation, or transduction (through viruses).

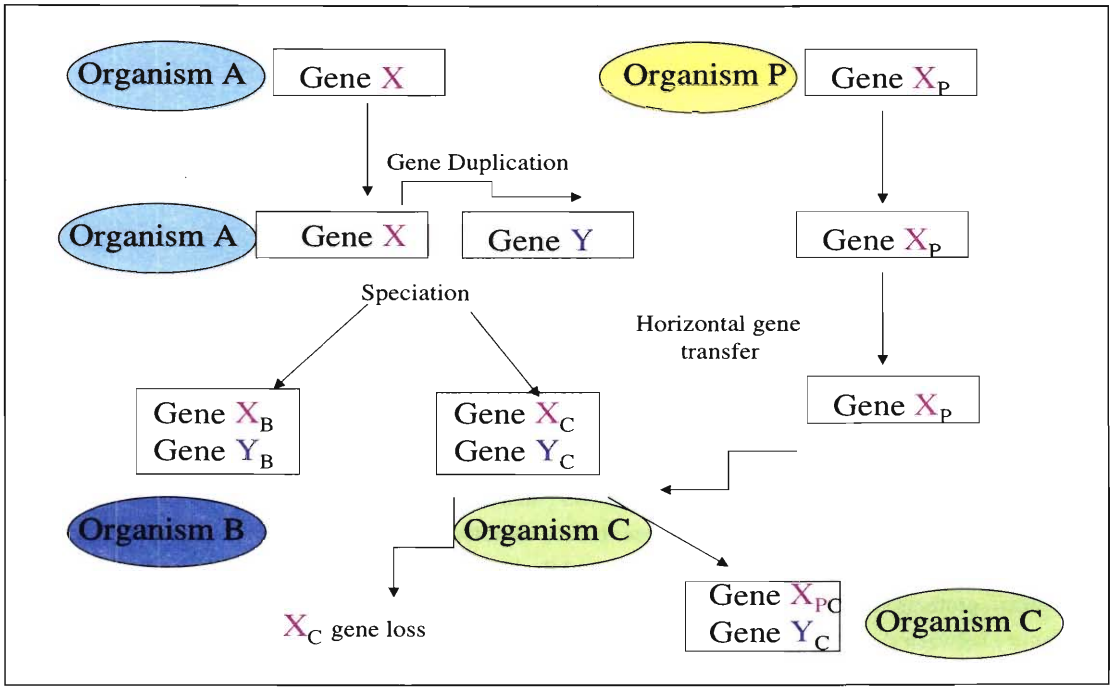


Figure 2. Schematic representation of the differences between orthologues, paralogues and xenologues, proteins that are homologous to each other but have yet originated a little differently. X and Y are paralogues; X_B and X_C are orthologues while XP_C is a xenologue, Y_B and Y_C are orthologues. X_B and Y_C as well as Y_B and X_C are also paralogues.

The Problem of Orthologue Identification

Most comparisons of genomes usually begin from the comparisons of the protein-encoding genes of two or more organisms that allows one to determine which pathways are absent or present in an organism. An important, and oftentimes difficult aspect, however, is to find out which proteins of the two organisms actually correspond to each other and are 'functionally equivalent'. On the face of it, this hardly appears to be a difficult problem, since the most similar proteins would be the most likely functional equivalents. In the majority of cases, this is in fact the case. However, difficulties arise from the variable rates of evolution of different proteins that result from differing environmental pressures faced by the organisms, and the presence of gene families that result from gene duplication events. Added to this is the problem that a significant fraction of

proteins have 'analogous' proteins (non-orthologous proteins) from other organisms (which have no sequence similarity) as their functional equivalents, having arisen from 'convergent evolution' rather than 'divergent evolution'.

This process of finding "which genes correspond to which" in the compared organisms, is really an attempt at 'orthologue identification' in the compared organisms. By stating that genes X and Y are orthologues, the underlying meaning is that these genes are playing the same role since they correspond to one another and share an evolutionary history.

By stating that genes X and Y are orthologues, the underlying meaning is that these genes are playing the same role since they correspond to one another and share an evolutionary history.

Why is orthologue identification so important? There are possibly three main reasons why this is such an important issue in comparative genomics. Firstly, orthologue identification allows reliable prediction of gene function in newly sequenced genomes. Second, in phylogenetic analysis, trees can only be constructed within sets of orthologues. Third, a complete list of orthologues is necessary for a meaningful comparison of genes and genome organization between organisms.

In general, two strategies may be applied in orthologue identification based on similarity: (i) In the first approach a statistical parameter is used to determine whether the sequence match is significant (using some pre-defined cut-off values). Most often the widely used algorithm for sequence similarity searches, BLAST (Basic Local Alignment Search Tool) is used and the BLAST '*E* value' used for determining cut-offs. The '*E* value' is a statistical parameter reflecting the probability of finding a similar sequence in the database. The *E* value takes into consideration the size of proteins, as well as databases being searched. This statistical parameter is very powerful, and also the most used, owing to it's amenability to automated analyses. However, employing arbitrary cut-offs does not take into account the fact that not all proteins evolve at the same rates. Many proteins evolve rapidly, while others evolve slowly. Furthermore, the same protein may evolve slowly in one organism, and rapidly in another organism depending on the evolutionary pressures faced

by the organisms. Eventually, one has to make a trade-off between sensitivity (the ability to detect even remote orthologues) and accuracy (the ability to accurately predict an orthologue) Using pre-determined cut-offs, one is likely to eliminate rapidly evolving proteins (*Figure* 3), or else one is left dealing with proteins that have common domains, but are not true orthologues to each other (*Figure* 4). Thus using the statistical parameters, additional parameters are also often imposed. As an example, proteins should align along 60-80% of their lengths. This constraint permits one to identify true orthologues as opposed to proteins that simply contain conserved motifs or domains. Using these criteria, in an analysis of orthologous proteins in the yeast, fly and the worm, different numbers of orthologous proteins were obtained and are shown in *Table* 2. The significant difference in these numbers perhaps most acutely defines the difficulty in identification of orthologous proteins in organisms. In a sense this is a major stumbling block to rapid automated comparative analysis of genomes.

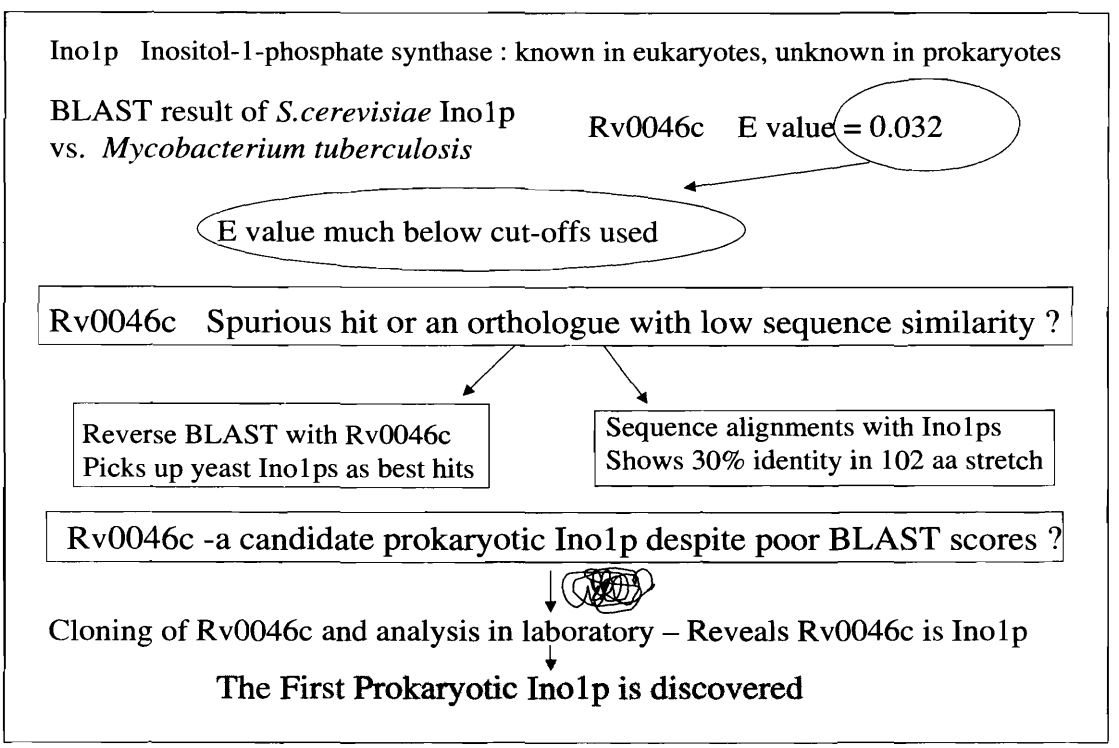


Figure 3. The difficulties in finding orthologues: The case of the prokaryotic Ino1p. Schematic representation depicting the discovery of the first prokaryotic Ino1p enzyme, an enzyme earlier thought to be restricted to eukaryotes. This discovery, carried out by Nandita Bachhawat and Shekhar Mande at IMTECH, is an example how an orthologue identification was made despite the very insignificant BLAST E values.

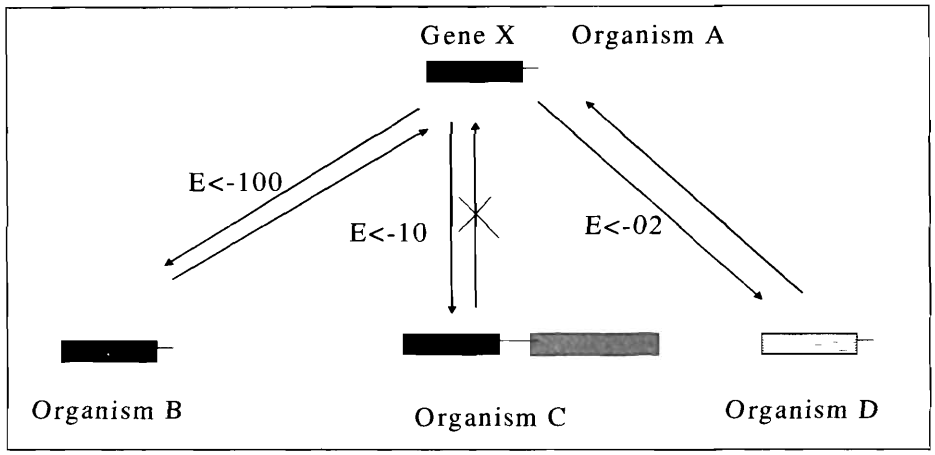


Figure 4. The 'domain' problem in orthologue identification. Schematic representation depicting how significant BLAST E values might give rise to spurious orthologue identifications.

In the second strategy, a non-statistical approach is taken where orthologous proteins are identified as those with the best matches to each other irrespective of the cut-off *E* values (or any other statistical parameter). This helps to overcome the problem of proteins having variable rates of evolution, and dispenses with a single statistical cut-off. This Best Hit Analysis method has been very successfully employed in the "COG database" (Clusters of Orthologous Groups). This approach accommodates vastly differing evolution rates of different genes and identifies closest homologues even if not statistically significant in itself. The COG approach was able to successfully predict the function of 100 proteins of *Helicobacter* that were not previously predicted by using merely the statistical parameter-based approach described above. However, this approach also needs to be employed with caution, since paralogous proteins are collapsed into a single orthologous cluster for the analysis.

Table 2.

Numbers of orthologous proteins between *Drosophila melanogaster* (fly) and *Saccharomyces cerevisiae* (yeast) at different BLAST *E* cut-off values. (Source: Chervitz *et.al.*, 1998)

Organisms	$E<10^{-10}$	$E<10^{-20}$	$E<10^{-50}$	$E<10^{-100}$	Additional constraints applied
Fly –Yeast	2345	1877	1036	433	Proteins should align over 80% of their length
Fly–Yeast	3986	2677	1266	504	Nil

An alternative to employing sequence similarity using E value cut-offs or clustering followed by best-hit, is to use phylogenetics based predictions of orthologues by 'phylogenomics'.

Despite their utility these methods are unable to distinguish paralogues from orthologues (wherever more than one homologue is present in a particular organism). An alternative to employing sequence similarity using E value cut-offs or clustering followed by best-hit, is to use phylogenetics based predictions of orthologues by 'phylogenomics'. In this approach, for a particular gene family, functions of homologues are overlaid onto a phylogenetic gene tree to infer functions of genes of interest.

One final point. Although the general assumption is that orthologues in different species retain their function, there is also a possibility that changes in function could also occur upon speciation. This may especially occur in distantly related species where the orthologous proteins may have diverged to a slightly different function.

"Phylogenetic Footprinting" and the Extraction of Information from Non-Coding Sequences

The nucleotide sequence of the genes encoding proteins, their organization or gene order (synteny) can also help in determining the correspondence of the different proteins in different organisms. In many cases, especially in higher eukaryotes where genes are 'buried' in introns, this has also helped to discover new genes. However, gene order is preserved only in organisms that are phylogenetically close. As these distances increase there are more opportunities for molecular processes of evolution: mutations, deletions, inversion, rearrangements to ultimately destroy the gene order completely. Even among the prokaryotes, gene order is not conserved except in closely related species. For example, gene order is conserved to only a limited extent between the gram-negative bacteria *Escherichia coli* and *Haemophilus influenze* but it is conserved in closely related bacterial species *Mycoplasma genitalium* and *Mycoplasma pneuomoniae*. Among the higher eukaryotes, gene order has been found to be preserved significantly in mouse and humans and becomes a useful aid for prediction of coding regions and genes. We might like to think

The nucleotide sequence of the genes encoding proteins, their organization or gene order (synteny) can also help in determining the correspondence of the different proteins in different organisms.

Organism Pairs	Phylogenetic Distance *	Gene Order Conservation	Coding Sequence information conservation	Non-coding information conservation
Saccharomyces cerevisiae –Schizosaccharomyces pombe	330MY	No	Yes	No
Human–mouse	75–80 MY	Yes	Yes	Yes
Human–chimpanzee	5–8 MY	Yes	Yes	Yes
Haemophilus influenzae– Escherichia coli	160 MY	Limited	Yes	No
Saccharomyces cerevisiae –Saccharomyces bayanus	5–20 MY	Yes	Yes	Yes
Escherichia coli– Shigella flexneri	25 MY	Yes	Yes	Yes
Escherichia coli K12 MG1655- *Escherichia coli* K12 W3110	40Y	Yes	Yes	Yes

*MY = Millions of Years

Table 3. Organism pairs, their phylogenetic distances and the conservation of information.

otherwise, but we are phylogenetically closer to a mouse than the gram negative bacteria *Haemophilus influenzae* and *Escherichia coli* are to each other!

Non-coding sequences, such as intergenic regions, promoters, terminators or introns, are far less conserved than protein-coding sequences, and have therefore posed a far greater challenge to computational methods to be able to discern important elements within these regions. Promoter elements for example, contain sequence motifs that are far shorter than genes, have greater degeneracy in the permissible sequence and can act at variable distances from the start site and in either orientation.

If one compares the non-coding sequences of organisms that are phylogenetically related in which the gene-order or synteny is

Promoter comparisons and analysis have been rewarding in yielding valuable new information on conserved motifs in these non-coding regions that has otherwise been very difficult to obtain by bioinformatic approaches and pattern finding programs.

conserved, the possibility that the non-coding sequences may have diverged only marginally and that important elements in these non-coding regions is still preserved would perhaps allow one to identify such elements. This strategy, now referred to as 'phylogenetic footprinting' was in fact able to identify, in a human–mouse comparison of the interleukin-4,-5 and -13 genes, a conserved sequence of 401 bp that was found to play a vital role in interleukin regulation. This sequence was in fact missed by earlier studies on the regulation of these genes through conventional approaches in wet-lab experiments.

In the yeast *S.cerevisiae*, this approach has been investigated and examined in a very exhaustive manner by the groups led by Mark Johnston and Eric Lander, by sequencing the genomes of the *Saccharomyces* genus (*S.mikatae*, *S.paraoduxus*, *S.bayanus*, *S.cariocanus*, *S.kudrivezii*, *S.castelli* and *S.unisoporus*). Comparative genomics of these organisms has led to several new insights into the coding as well as the non-coding regions of the *S.cerevisiae* genome. From an exhaustive comparative analysis, over 500 Open Reading Frames (ORFs) were found to be incorrectly described as genes (mis-annotated) and could be eliminated as a result of this analysis. In addition 43 new genes of size less than 100 amino acids and many new introns in many existing genes were also predicted. Furthermore, protein start site annotations (the correct 'ATG') were corrected in 72 ORFs.

Promoter comparisons and analysis have been rewarding in yielding valuable new information on conserved motifs in these non-coding regions that has otherwise been very difficult to obtain by bioinformatic approaches and pattern finding programs. By finding such patterns, the corresponding genes have been grouped on the basis of possible similarity in regulation, and many unknown ORFs could be predicted to function in distinct pathways. A database containing conserved elements in orthologous promoters across these organisms was recently created by us to facilitate investigations on these cis-regulatory elements.

The phylogenetic footprinting approach, which was first systematically done for the yeast *Saccharomyces cerevisiae* has also been initated for other organisms, like *Drospophila melanogaster*, *Arabidopsis* and in humans where comparison with other primate sequences may yield even greater information than that of human-mouse. The analysis of very closely related species (such as humans and other primates) has been termed 'phylogenetic shadowing' but is based upon the same rationale. These methods have been a major boost to experimental researchers in uncovering information of non-coding DNA, where other strategies are either largely unsuccessful, or simply impractical.

One of the biggest surprises from genome sequencing projects and comparative genome analyses has been the relatively high level of horizontal gene transfer seen in different microbes.

Insights into Genome Fluxes and the Processes of Evolution

The compositions of the genomes of two or more organisms also reveals a great deal of insights about different aspects of the evolution of the genome of the organism, genome dynamics such as gene (or pathway) loss or gene duplications and horizontal gene transfer (HGT). This analysis is vital for understanding evolutionary biology and the mechanistic principles involved in the different evolutionary processes. This has now become possible with the availability of a large number of genomes.

From an evolutionary biology perspective, whole genome comparisons provide molecular insights into the processes of evolution that include the molecular events responsible for the variations and fluxes that occur through a genome. These include processes like inversions, translocations, deletions, duplications and insertions.

One of the biggest surprises from genome sequencing projects and comparative genome analyses has been the relatively high level of horizontal gene transfer seen in different microbes. Microorganisms display upto about 10-11% of their genes as those that have arrived from HGT. In *E.coli* this has been estimated to be about 18%. This extensive intermixing of genomes across organisms has prompted many authors to suggest

While horizontal gene transfer tends to increase the genome size, a complimentary movement is that of reductive evolution that results in loss of genes through mutation, decay and deletion from an organism.

that a microbe is not merely an organism but a 'global superorganism'

How can we know if a gene or a group of genes has resulted from horizontal gene transfer? The unambiguous identification of HGT is often difficult. Several features are combined to establish if a gene (or group of genes) has been transferred through HGT. The first is nucleotide composition (GC-content). These could be either simple compositions or more complex patterns such as dinucleotide or tri-nucleotide patterns. Being relatively constant, it takes time for genes transferred from other organisms to become closer to the host genome. However the problem here is that some regions have selection/mutational bias, while in other cases transfers may be with organisms with similar composition. Differences in codon usage, variations in gene density, and examination of the phylogenetic positions of proteins are additional means to detect HGT.

While horizontal gene transfer tends to increase the genome size, a complimentary movement is that of reductive evolution that results in loss of genes through mutation, decay and deletion from an organism. This results from the adaptation and evolution of the organism to a different environmental niche making the functions of some genes redundant. One of the most striking examples of this phenomenon has been revealed in the genome comparisons of obligate intracellular parasites and endosymbionts. These include the mammalian parasites *Rickettsiae*, *Chlamydia* and the endosymbiont *Buchnera* which resides within certain aphids. These prokaryotic organisms have a much reduced genome size (~1 MB) in comparison to most other free living bacteria (~4 MB), and many genes have been lost in these organisms while only a few have been gained from the host organism. The obligate intracellular parasite, *Mycobacterium leprae* has a slightly higher genome size (~3.3 MB) than either Rickettsiae or Chlamydia. However comparisons with *M.tuberculosis* (~4.4 MB) have revealed that 27% of the genome of *M.leprae* contains pseudogenes while another ~24% is non-coding and may consist largely of pseudogenes that have extensively mu-

tated beyond recognition and are on their way out. A similar, relatively high content (~24%) of non-coding sequences was observed in Rickettsiae and were found to be primarily pseudogenes that are in the process of being lost. These comparisons suggest that when one observes a genome sequence of an organism it is really a 'snapshot in evolutionary time and space'.

When one observes a genome sequence of an organism it is really a 'snapshot in evolutionary time and space'.

The process of gene loss is however not restricted only to parasites or symbionts living within a host. In comparing the extent of gene loss of *Saccharomyces cerevisiae* since the common ancestor with the fission yeast *Schizosaccharomyces pombe* it was found that approximately 300 genes were lost and an equal number had diverged exceedingly indicating a 10% loss and divergence. In many cases these losses were confined to important groups such as those involved in pre-mRNA splicing, RNA modification, post-transcriptional gene silencing, nuclear structure maintenance and protein-folding processes. In *S.cerevisiae* also, the process has not become static, since current estimates indicate that about 3% of the *S.cerevisiae* genome has pseudogenes, and indications are that gene loss is an ongoing process in the life of an organism.

A third aspect of genome evolution that has been revealed from genome comparisons is the proliferation of gene families by gene duplication followed by divergence, especially when one examines the movement of organisms upward, for example from lower unicellular eukaryotes to that of multicellular eukaryotes.

The Impact of Comparative Genomics in Phylogenetic Analysis:

Most phylogenetic reconstructions and evolutionary tree-building have been essentially using data from single genes. The analysis is either based on the small subunit rRNA (16S rRNA/ 18SrRNA) or on a single, important, and highly conserved protein which has complex interactions with many other RNAs and proteins (making conservation of the protein almost certain). These proteins include the highly conserved RNA

Most phylogenetic reconstructions and evolutionary tree-building have been essentially using data from single genes.

In addition to horizontal gene transfer, however, there are other reasons that are responsible for these discrepancies seen in phylogenetic trees.

polymerases and elongation factors present in all kingdoms of life. However, often differences are seen between the protein phylogenic trees and those constructed with the 16s rRNA.

One of the reasons for the anomalies is due to the extensive horizontal gene transfer that genome comparisons have revealed. In addition to horizontal gene transfer, however, there are other reasons that are responsible for these discrepancies seen in phylogenetic trees. This involves the presence of ambiguous paralogues, variable rates of evolution of different proteins, methodological problems such as long branch attraction and mutation saturation. Furthermore, one gene corresponds to only a tiny fraction of the genomic material. For eg., 1.8 kb of 16s rRNA~0.05 % of a microbial genome (4 MB). Therefore, focusing on single genes ignores the bulk of the genomic material. Whole genome comparisons could lead to much better constructions of phylogeny trees. Many new whole genome approaches are also being attempted. These could be either alignment-free genome trees based on some 'word frequency' such as dinucleotide comparisons or frequency of occurrence of certain amino acid stretches or even amino acid compositions of proteins. Or else they could be based on gene order (for closely related organisms), average gene sequence similarities ('blastology' methods), shared gene content, or 'phylogenomic' methods where traditional gene trees are combined to yield 'supertrees'.

An example of how genome comparisons can assist in resolving controversies is in the phylogenetic position of the mammalian parasite, the microsporidion, *Encephelitozoan cuniculi*. These parasites lack mitochondria and early evolutionary trees based on the small subunit rRNA have considered it to be a deep-branching eukaryote, a primitive eukaryote that evolved prior to the evolution of the mitochondria. However the genome sequence now clearly reveals the presence of several mitochondrial genes within the parasite genome. Analysis and comparison of genomes indicate it to have a clear fungal origin, suggesting that the organism be placed on the phylogenetic tree closer to fungi and subsequent to the evolution of the mitochondrion (*Figure* 5).

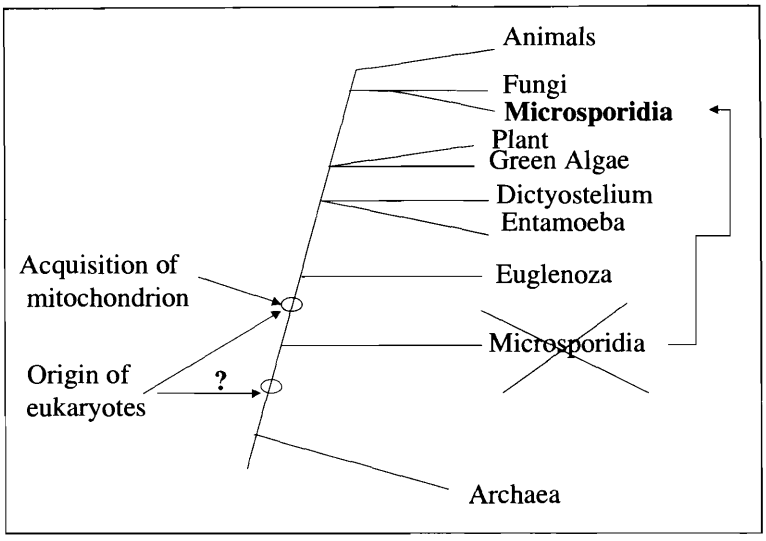


Figure 5. Schematic depiction of Microsporidia's phylogenetic position based on Small Subunit RNA (SSU rRNA) as an early branching eukaryote that evolved prior to the acquisiton of mitochondria, and it's subsequent placement based on a composite gene phylogeny where it was placed closer to fungi. The latter placement has been confirmed by the complete sequence of the micro-sporidia, Encephalitozoon cuniculi, where despite the absence of mitochondria, the presence of several mitochondrial genes could be observed.

Comparative Genomics in Drug Discovery

Comparative genomic studies throw important light on the pathogenesis of organisms, throwing up opportunities for therapeutic intervention as well as help in understanding and identifying disease genes.

By comparison of the human-mouse genomic sequences, a 5th apolipoprotein gene, APOA5, was discovered adjacent to the APOA4 in the region of the genome that contained the clusters of the apolipoprotein genes. The authenticity of APOA5 was subsequently investigated in knockout mice and found to encode a functional apoliprotein that in fact played an important role in determining the plasma triglyceride levels and has now been implicated in several disease conditions.

One of the most important fallouts of comparative analyses at a genome-wide scale is in the ability to identify and develop novel drug targets. Thus, if one is looking for antibacterial, antifungal, or antiprotozoal proteins to be used as targets, comparative genome analysis can reveal virulence genes, uncharacterized essential genes, species-specific genes, organism-specific genes, while ensuring that the chosen genes have no homologues in humans (*Figure* 6). With this background information it allows

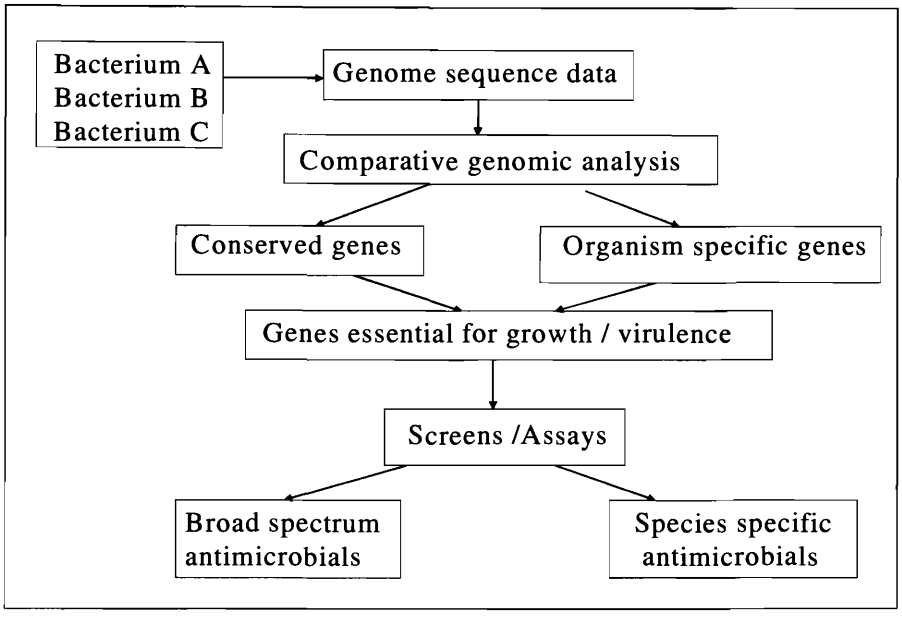


Figure 6. Comparative genomics in drug discovery programs. A flow chart diagram explaining how comparative genomics can facilitate drug discovery programs for the discovery of new antimicrobials.

researchers to design more rational experimental strategies to restrict the number of potential drug targets to those with most likely chances of success. Although the *in silico* analysis needs also to be validated by experimental approaches, a great deal of time and money is saved by limiting the number of targets to be studied by this prior analysis.

In an alternative approach, unique or analogous enzymes may be investigated as drug targets. Thus although a pathway may be conserved between the two organisms, the presence of an analogous enzyme at one step allows one to target that particular step in the pathway since inhibitors of this enzyme (in pathogens for example) would not affect the analogous enzyme in the host organisms (such as humans). One surprising observation from comparative genomics studies is the high number of such 'non-orthologous gene displacements' where proteins lack orthologues in other organisms, even though the enzymatic activity is present. In fact when one compares two bacteria such as *Mycoplasma genitalium* and *Haemophilus influenzae* approximately 10% of the total orthologues are in fact a case of non-orthologous gene displacement.

One surprising observation from comparative genomics studies is the high number of 'non-orthologous gene displacements' where proteins lack orthologues in other organisms, even though the enzymatic activity is present.

Looking Beyond Comparative Genomics

As we progress with leaps and bounds in understanding the flux and dynamics of evolution through comparative genomics, the immediate question is what would be the next level of comparisons? Comparisons of genomes, while it has provided a great deal of information on the dynamics of evolutionary processes, still leaves us with a somewhat 'static picture' of the organism and the cellular machinery. Perhaps, therefore, the next level of comparisons would be a more dynamic one, where one compares pathways and fluxes and movements of proteins and molecules within the cell, and subsequently between cells and organisms. We also know that in higher eukarytoes there are more transcripts (mRNA) than genes through the mechanisms of alternative splicing, and therefore 'transciptome comparisons' become increasingly more important. These would also include how an organism responds to different conditions, looking at its senstivities and robustness. Similar questions that have been raised have given birth to a new branch called Systems Biology. Comparisons at this level would provide us with a more dynamic picture of the cell.

As we progress with leaps and bounds in understanding the flux and dynamics of evolution through comparative genomics, the immediate question is what would be the next level of comparisons?

Prokaryote organisms have a very small percentage of non-coding junk DNA (less than 15% of the genome). This is in contrast to almost 99% of the DNA being relegated as 'junk DNA' in higher eukaryotes. An important issue is the possible role if any of all this junk DNA One hopes that possibly comparative genomics, which is proving to be so powerful in the understanding of non-coding elements might finally be able to reveal what all this junk DNA is really about.

As comparative genomics moves from between kingdoms to between genus to between species analysis, the next step is to carry out comparisions between individuals or strains that are members of a particular species. This would allow us to investigate variations at the individual level and to enable one to determine the propensity of an individual to respond to a drug or to come down with a disease or infection. Would it also enable

one to artificially construct a prokaryotic or eukaryotic genome with a minimal number of genes?

There is perhaps yet another tormenting question – one that philosophers have been debating for centuries – that one believes would be addressed more aggressively. And that is the question of the increase in the cognitive powers of humans. The chimp is evolutionarily so close to the human (~99% similar in sequence) and yet there seems to be a major differnce between the chimp and the human in terms of the ability to 'conceptualize'. What networks and pathways might have led to this big leap? We hope that this question will be answered soon.

The excitement is only just beginning.

Suggested Reading

[1] C A Ball and M P Cherry, Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms, *Curr. Opin. Chem. Biol.*, Vol.5, p. 86, 2001.

[2] Chervitz *et.al*, Comparison of the complete protein sets of worm and yeast: Orthology and divergence, *Science*, Vol.282, p.2022, 1998.

[3] G M Rubin, *et al*, Comparitive genomics of the eukaryotes, *Science*, Vol.287, p. 2204, 2000.

[4] M Y Galperin, D R Walker and E V Koonin, Analogous enzymes: Independent inventions in enzyme evolution, *Genome Res.*, Vol. 8,pp. 779-790, 1998.

[5] M Y Galperin and E V Koonin, Searching for drug targets in microbial genomes, *Current Opinions in Microbial Genomes*, Vol.10, No.6, pp.571-578, 1999.

[6] J P Gogarten and L Olendzenski Orthologs, Oaralogs and genome comparisons, *Curr. Opin. Genetics & Develop.*, Vol.9, pp.630-636, 1999.

[7] M A Nobrega and L A Pennachio, Comparative genomics as a tool for biological discovery, *J.Physiol.*, Vol.554,pp. 31-39, 2003.

[8] M Kellis, N Patterson, M Endrizzi, B Birren, E S Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature*, Vol. 423, pp. 241-254, 2003.

[9] P Cliften, P Sudarsanam, A Desikan, L Fulton, B Fulton, J Majors, R Waterston, B A Cohen, M Johnston M, Finding functional features in Saccharomyces genomes by phylogenetic footprinting, *Science*, Vol. 301, pp. 71-76, 2003.

Address for Correspondence
Anand K Bachhawat
Institute of Microbial Technology
Chandigarh 160 036, India
Email:anand@imtech.res.in