Sequence Evolution :

Evolutionary Concept In Genomics

Presented By:

Imeh Akpan (s183179)

Why does the sequence of a ribosomal protein from different species exhibit considerable diversity and yet they align unequivocally?

INTRODUCTION

- In the past, gathering information about your potential marriage or business partner would have started with the simplest questions as "What family does he or she come from?"

- Affiliation with some family ties would have immediately started pointing you in the direction for further enquiries - a general ideal of what might be expected from some certain individuals.

- The same approach is used in predicting potential functions for a newly sequenced gene and its protein product.

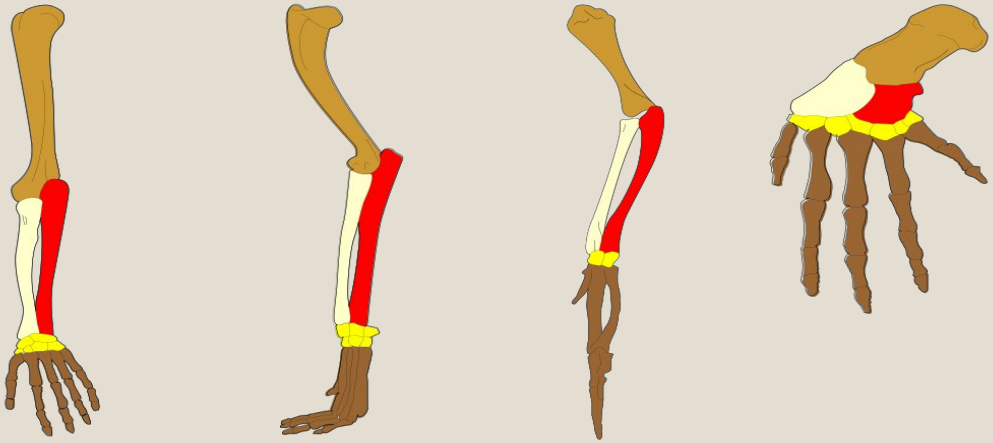
The main idea:

- Sequence analysis aims at finding important sequence similarities that would allow one to infer **homology**. Since the mid-19th century, zoologists and botanists have learned to make a distinction between homologous organs and similar (analogous) organs.

cont..

- This simple concept tends to get extremely complicated when applied to protein and DNA sequences. "Sequence homology" has been assigned as a keyword to more than 80,000 papers in MEDLINE.

Homologous Organs



- "homology" is used basically as a glorified substitute for "sequence (or structural) similarity".

What is Homology

- The term "homology" has been used often to designate quantifiable similarity between sequences.

- We believe the notion of homology is of major fundamental and practical importance. In my opinion, misuse of the term 'homology' has the potential of washing out the meaning.

- A real problem arises only when the **similarity between two given sequences is much lower**, so it is not immediately clear how to properly align them.

- Lower levels of similarity might be indicative of homology provided that one or more of the following criterias applies:

Criterias:

- (i) the similarity extends over a long stretch of sequence and is statistically significant by criteria known to be reliable (such as in BLAST algorithm and its derivatives);

- (ii) although the sequence similarity is low, the same pattern of identical and similar amino acid residues is seen in multiple sequences; or

- (iii) the pattern of sequence similarity reflects the similarity between experimentally determined structures of the respective proteins or at least corresponds to the known key elements of one such structure.
-

Model of Sequence Evolution

- In the rest of this presentation, I want to provide answer to the question "why is sequence and structural similarity considered to be evidence of homology (common origin) . in the first place?

- Once we are confident that a particular similarity is not false, but rather, according to the above criteria; it represents certain biological reality.

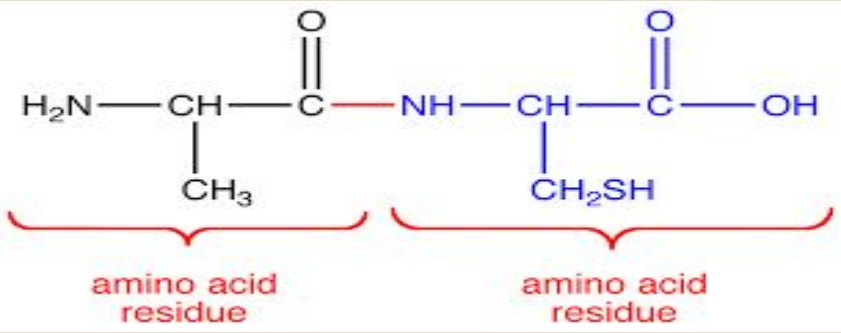
Again, is common ancestry the only explanation?:

- The answer is: no, a logically consistent alternative does exist and involves **convergence** from unrelated sequences.

Convergence Hypothesis

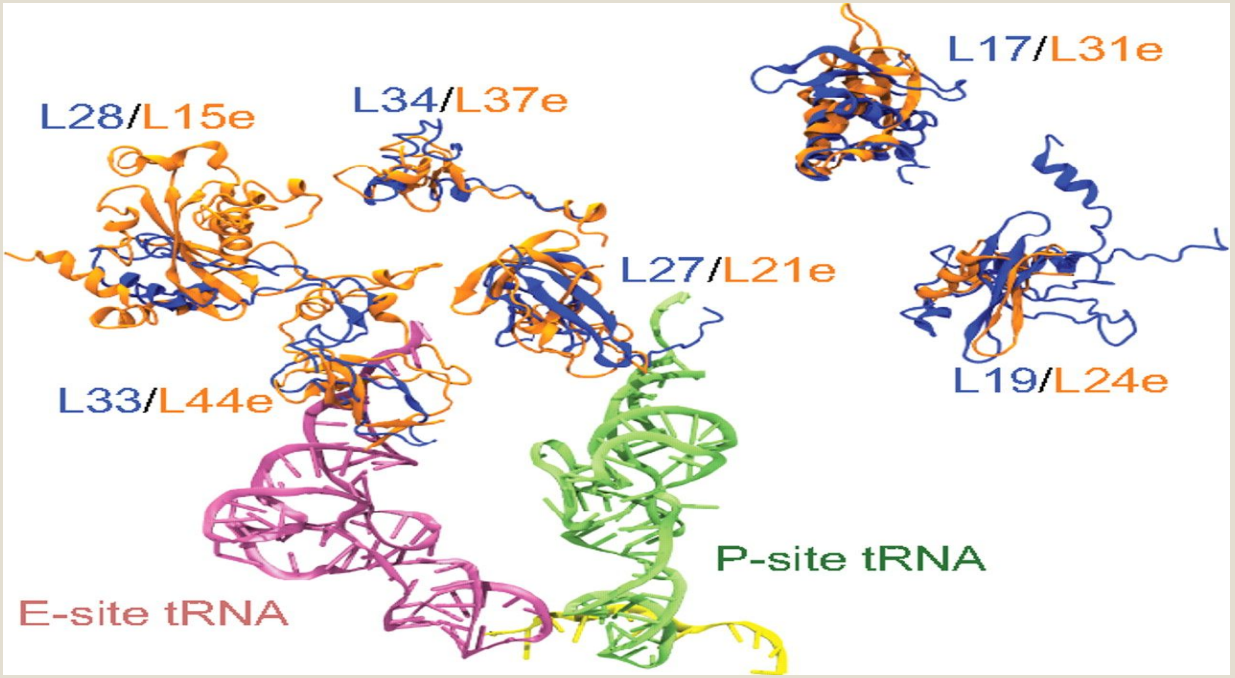
They say that sequence and structural similarities between proteins are observed because the shared features are strictly required for these proteins to perform their identical or similar functions.

Indeed, if we take two sequences of 100 amino acid residues each that have, say, 80% identical residues, we can calculate the probability of this occurring by chance, find that it is so low that such an event is extremely unlikely to have happened in the last 5 billion years, and conclude that the sequences in question must be homologous (share a common ancestry)



cont..

For example, although sequences of the ribosomal protein L36 from different species exhibit considerable diversity and only a single amino acid residue is conserved in all the sequences, they align unequivocally and are indisputable homologs



The Neutral Theory

A pan-adaptationist view of evolution would hold that functional convergence is the sole (or at least the principal) factor responsible for similarity between proteins

Although there are a few compelling arguments against this convergence hypothesis but to formally disprove this paradigm might not be possible.

Putting all together, convergence hypothesis is equivalent to the statement that most, if not all, amino acid residues in a protein are fixed through positive selection. This runs against the neutral theory of molecular evolution, which has shown that, given the known parameters of animal populations, positive selection could not be responsible for the majority of amino acid substitutions, which are therefore effectively neutral.

An obviously contrasting models of Evolution

Divergence Evolution

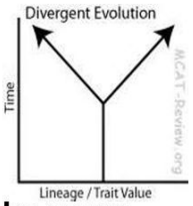
- The second, probably most convincing, argument against convergence as the principal explanation for the observed similarities between proteins has to do with the nature of structural constraints associated with a particular function.

- A fundamental observation is that a single function, such as catalysis of a specific enzymatic reaction, is often performed by two or more proteins that have unrelated structures.

- But by inference, we are justified to conclude that whenever statistically significant sequence or structural similarity between proteins or protein domains is observed, this is an indication of their **divergent evolution** from a common ancestor or, in other words, evidence of homology.

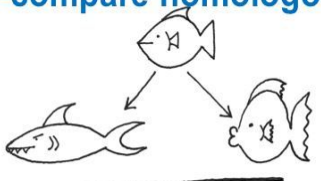
Types of Sequence Evolution

TYPES OF EVOLUTION

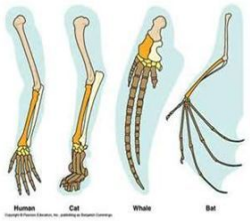


Divergent evolution: populations become more and *more dissimilar* to adapt to the environment

-compare homologous structures

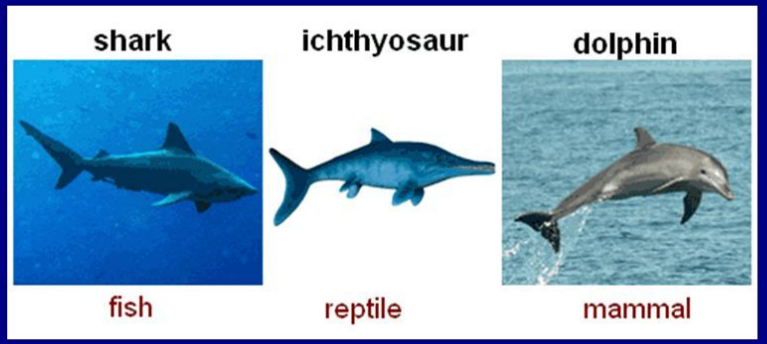


-one species evolves into two different species



Convergent Evolution

- **Different species become more similar.**
  - Example #1: Birds and bats
  - Example #2: Dolphins and sharks

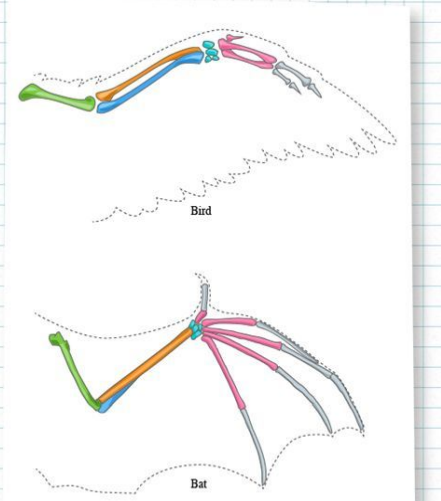


cont..

Homologies

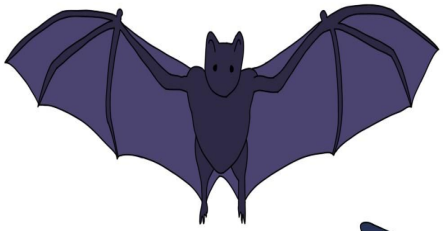
sapling learning

- Anatomical homology
  - Result of common ancestor with that trait
- Convergent evolution
  - Creates similar structures/functions
  - Are not anatomical homologies



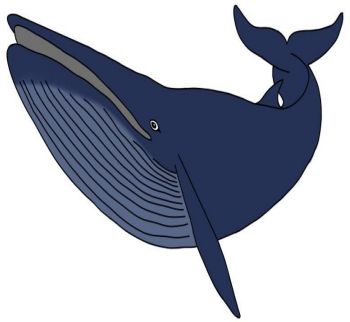
CONVERGENT EVOLUTION OF ECHOLOCATION

Both bats and whales use echolocation to find food



bats hunt for food at night and use echolocation to "see"

whales have limited access to sunlight in the ocean and use echolocation to find their way around



Conclusion

- it should be emphasized that demonstration of homology is central to the interpretation of similarities between proteins.

- The feasibility of this conclusion, which sometimes is reached on the basis of limited similarity, is what makes sequence and structure comparison the major staples of computational biology and inspires the development of increasingly sensitive methods for such comparisons.

- Indeed, under the notion of homology, a sequence or structural alignment becomes a powerful tool for evolutionary and functional inferences.

The Essence of Sequence Alignment

Once sequences are correctly aligned, homology implies that the corresponding residues in homologous proteins are also homologous, i.e. derived from the same ancestral residue and, typically, inherit its function. If the residue in question is the same in a set of homologous sequences, we say that it is (evolutionarily) conserved.

Thus, homology lends legitimacy to the transfer of functional information from experimentally characterized proteins (or nucleic acids) to uncharacterized homologs, the single most common and practically important application of computational methods in molecular biology. Conversely, an alignment of non-homologous sequences is inherently meaningless and potentially misleading.

cont..

Even if such an alignment attains a relatively high percentage of identity or similarity, no conclusions at all can be inferred from the (fake, in this case) correspondence between aligned residues.

This is why phrases like "significant homology" or "percent homology" are so absurd. Homology is a qualitative notion of common ancestry.

As long as homology is established, 10% identical residues between two protein sequences could be highly meaningful and amenable to functional interpretation. In contrast, even 30% identity between two sequences that are not homologous in reality could be totally misleading.

Homologs

- One of the main objectives of DNA and protein sequence analysis is to identify homologous sequences.
- **The two categories of homologs are orthologs and paralogs.**
- Orthologs are evolutionary counterparts derived from a single ancestral gene in the last common ancestor.
- Paralogs are genes that evolved through duplication within the same (perhaps ancestral) genome.
- These definitions were first introduced by Walter Fitch in 1970 and remained virtually unknown until the advent of genomics.

-b-

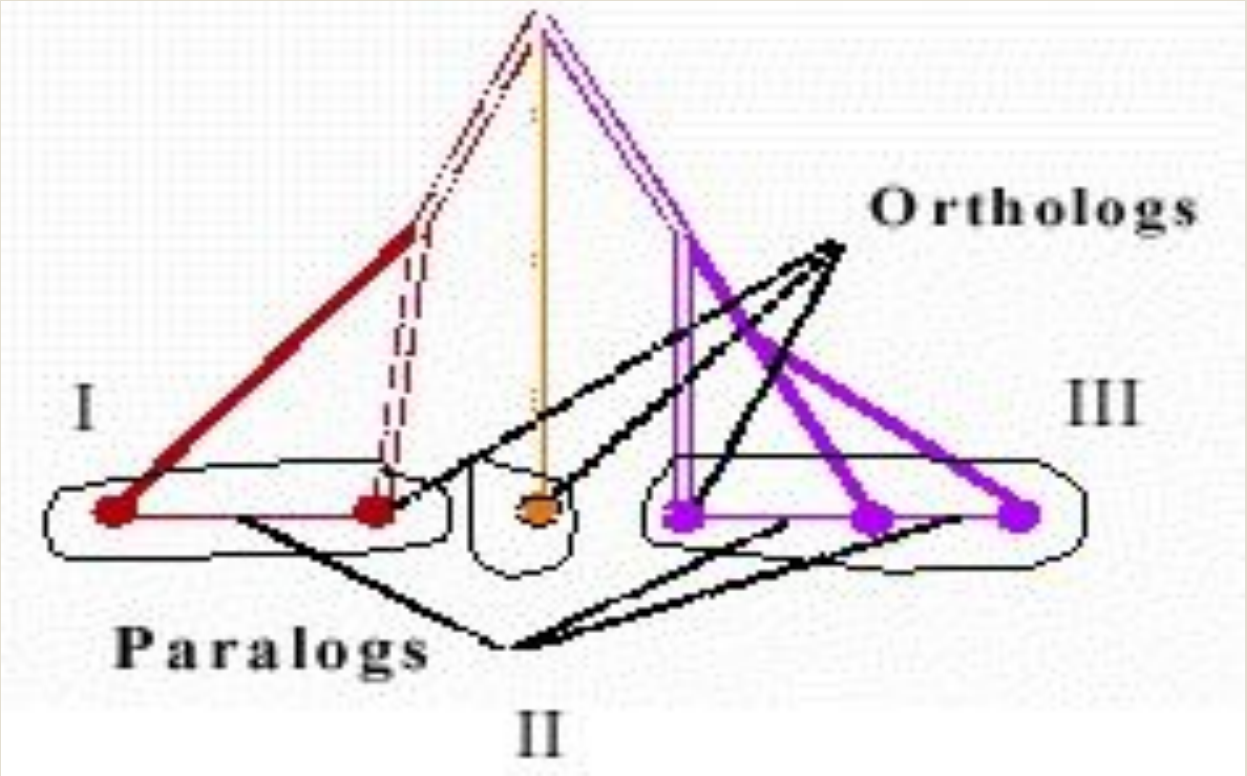
- In contrast, paralogs tend to evolve new functions, and study of paralogous families may provide means for understanding adaptation.
- Classic examples include animal olfactory receptors or nuclear hormone receptors, vast families in which an astonishing repertoire of specificities evolved.

- The interplay of speciation events, leading to the divergence of orthologs, and duplications, giving rise to paralogous families, results in complex evolutionary scenarios.

- The relationships between homologs could become tricky if some genes in certain lineages have been lost during evolution. In such cases, genes that, at face value, appear to be orthologous may actually be paralogs.

-b-

- Orthologous and paralogous genes in three lineages descending from a common ancestor. Gene sets I, II, and III should be considered co-orthologous.



-b-

- Identification of orthologs is only possible when complete sets of genes from two or more genomes are compared.

- In principle, complete phylogenetic analysis of all groups of homologous genes is required to decipher true orthologous relationships. "Shortcut" approaches have been developed to circumvent the need for comprehensive phylogenetic analyses.

Genome Evolution

- Although still a young discipline, comparative genomics has matured enough to allow delineation of the most common and important types of events that occur during genome evolution.

- These include different forms of genome rearrangement, gene duplication, and more specifically, lineage-specific expansion of gene families, lineage-specific gene loss, horizontal gene transfer, and non-orthologous gene displacement.

Evolution of gene order

- Comparing the first sequenced genomes showed that gene order is much less conserved than protein sequences. Genomes of the closely related bacteria Mycoplasma genitalium and M. pneumoniae were found to be scrambled. In the chlamydial genomes, a genome-scale alignment is readily traceable along the main diagonal, although gaps in the alignment and two major inversions are equally obvious.

- There is important conservation of gene order within operons, the units of prokaryotic gene coregulation. In each genome, 5% to 25% of the genes belong to conserved (predicted) operons. operons are strings of genes that are shared with at least one relatively distant genome. As should be expected, this fraction gradually increases as new genomes are sequenced.

-b-

E. coli K-12 has seven times more genes than the aphid symbiont Buchnera sp. Two more representatives of gamma-proteobacteria, H. influenzae and P. multocida, have 2.5 times fewer genes than E. coli. Baker's yeast S. cerevisiae, for example, has about 6,000 genes, 2,000 fewer than in its relatives, such as Aspergillus. Parasites might not need a complicated web of metabolic pathways as long as they can fetch nutrients from their host.

The vertebrate proteins in this list, including the uncharacterized ones, are obvious homologs of the goose lysozyme. The phage protein is more dissimilar and, in this case, the issue of homology is worth some investigation.

However, the sequence similarity between lysozymes and this phage protein is statistically significant and their multiple alignment shows a consistent pattern of shared residues, thus establishing homology

OBJECTIVE OF DNA SEQUENCING

An analysis of gene loss in bacterial parasites showed that, in many cases, it led to the elimination of entire pathways. A similar trend toward co-elimination of functionally connected groups of proteins, such as the signalosome and the spliceosome components, has been detected in yeast.

Lineage-specific expansion of gene families

Genome comparisons suggest that lineage-specific expansion of paralogous gene families is one of the major mechanisms of adaptation. In pathogens M. tuberculosis and H. pylori, the most conspicuous expansions are those of genes encoding factors involved in interactions with and survival within the host organisms.

Fig : Movement of Glowworm to their Local Optima

Expansion of signaling domains in C. elegans

In eukaryotes, lineage-specific expansion of certain protein families is even more evident than in prokaryotes. A comparison of the genome counts of signaling domains in the nematode C. elegans against the corresponding numbers

Lateral gene transfer refers to acquisition of genes from organisms that belong to other species, genera, or higher taxa. Mechanisms include conjugation, acquisition of plasmids, and viral (phage) infection. These events are common and do not stir much controversy. Long-range lateral gene transfer across taxa has been considered to be extremely rare and unimportant.

horizontal gene flow between closely related species turned out to be much more pervasive than ever suspected before. Lawrence and Ochman estimate that as much as 25% of the E. coli genome consists of recently acquired "foreign" genes. In the 100 million years since the split between Escherichia and Sal

Expansion of signaling domains in C. elegans

In addition, genome comparisons helped to uncover numerous cases of (predicted) horizontal gene transfer between organisms belonging to distinct phylogenetic lineages. Archaeal genomes presented a particularly striking picture, with some genes having close homologs only among eukaryotes. These observations could be explained by massive gene exchange between archaea and bacteria. This hypothesis was further supported by genome analysis of two hyperthermophilic bacteria, A. aeolicus and T. maritima.

We believe that the demonstration of the evolutionary prominence of lateral gene transfer can be considered the single greatest change in perspective in biology brought about by comparative genomics. A new round of controversy has been sparked by the discovery of genes of possible bacterial origin in the human genome [488]. In Chapter 6, we revisit this issue and discuss implications of large-scale lateral gene transfer for the "tree of life"

Non-orthologous gene displacement and the minimal gene set concept

Enzyme recruitment is a common evolutionary phenomenon leading to non-orthologous gene displacement. Enzyme recruitment seems to be particularly common in organisms that have adapted to novel ecological niches. Most of the enzymes that are responsible for the biosynthesis of polyketide antibiotics appear to be recent recruits from the enzymes of fatty acid biosynthesis. Perhaps the most remarkable example is the evolution of apyrase, the enzyme secreted by blood-sucking insects into the blood of human or other mammalian victims to prevent or slow down blood clotting.

We believe that the demonstration of the evolutionary prominence of lateral gene transfer can be considered the single greatest change in perspective in biology brought about by comparative genomics. A new round of controversy has been sparked by the discovery of genes of possible bacterial origin in the human genome [488]. In Chapter 6, we revisit this issue and discuss implications of large-scale lateral gene transfer

A scenario for the evolution of non-orthologous gene displacement via an ancestral redundancy stage and lineage-specific gene loss.

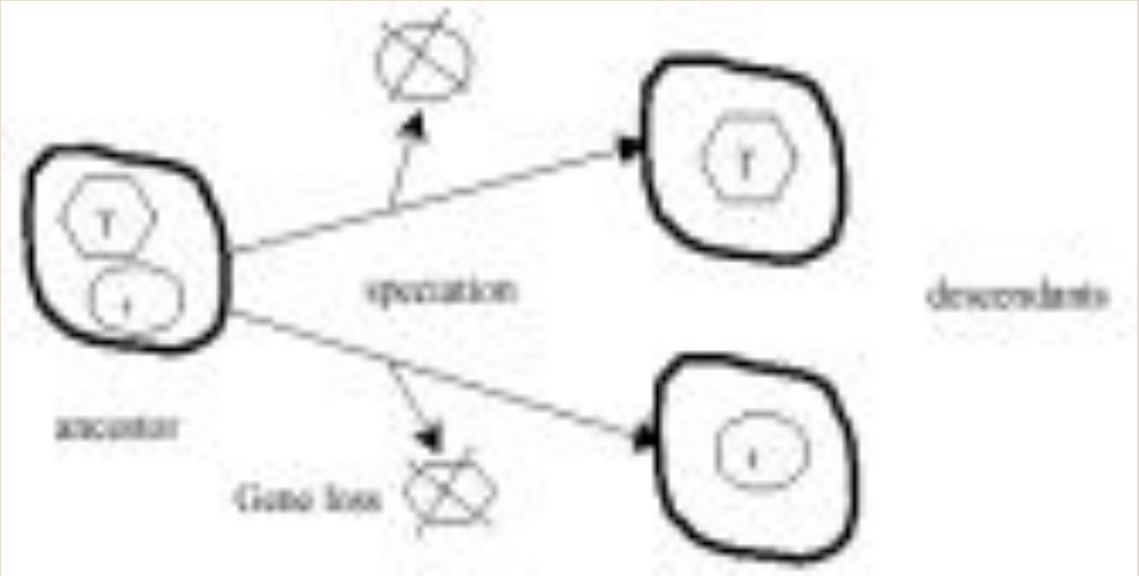


Fig : Movement of Glowworm to their Local Optima

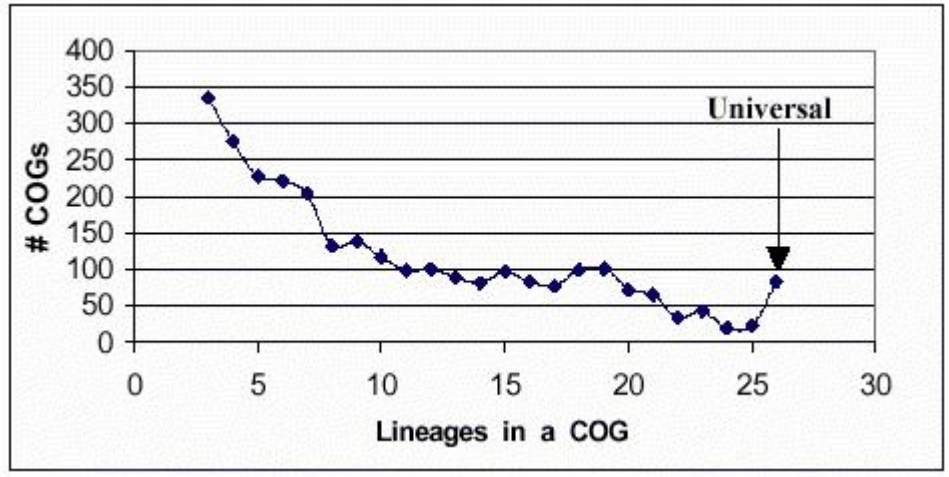
Non-orthologous gene displacement and the minimal gene set concept

Proteins responsible for the same function in different organisms typically show significant sequence and structural conservation. Only about 65 orthologous protein sets are universally represented in all sequenced genomes. This number is much lower than the number of essential functions, indicating that other such functions are performed by unrelated (or at least non-orthologous) proteins in different life forms. This major evolutionary phenomenon, which came to light already in the first comparisons of sequenced genome, was dubbed non-orthologous gene displacement.

In Chapter 7, we look at the comparative genomic of central metabolic pathways. We encounter numerous cases of non-orthologous gene displacement and, specifically, enzyme recruitment. It is worth noting that enzyme recruitment can be legitimately described as independent, convergent evolution of the same enzymatic activity. The idea ofnon-orthologistous gene displaced was originally developed in conjunction with the

Non-orthologous gene displacement and the minimal gene set concept

A subsequent large-scale experimental study has shown that most of the genes included in this theoretical minimal gene set were, indeed, essential in M. genitalium. sequencing of additional genomes and the corresponding genome comparisons have clearly shown that this early reconstruction vastly underestimated the extent of non-orthologous gene displacement. Only about 65 genes seem to be truly ubiquitous in cellular life forms, comprising perhaps 25% of the minimal set



Phyletic patterns (profiles)

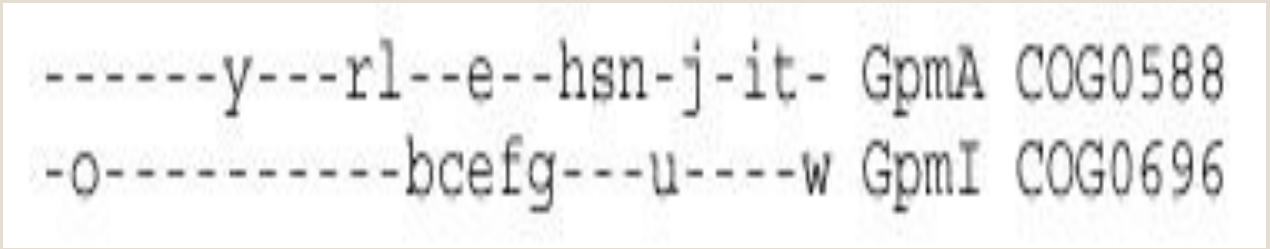
There is no necessity in yet another form of phosphoglycerate mutase, which has been designated GpmB in E. coli. This example shows the impressive power of the comparative-genomic approach for predicting gene functions. This methodology is discussed in greater detail later in this book.

```
---p--yqvdrlbcefghs--j---- GpmB COG0406
```

Only E. coli encodes both forms of the enzyme, whereas other organisms encode either one or the other. Several organisms do not encode either of the two forms of this enzyme. One might suggest that there should be an additional, third form of phosphoglycerate mutase, which is encoded in archaeal genomes and also in T. maritima, A. aeolicus, and D. radiodurans.

Phyletic patterns (profiles)

Most protein families show a "patchy" distribution among the sequenced genomes. Majority of COGs are represented in only three or four phylogenetic lineages. Phyletic patterns (profiles) show the presence or absence of a COG in each analyzed species. This approach provides a convenient way to compare genomes and investigate evolutionary history of cellular functions.



Only E. coli encodes both forms of the enzyme, whereas other organisms encode either one or the other. Several organisms do not encode either of the two forms of this enzyme. One might suggest that there should be an additional, third form of phosphoglycerate mutase, which is encoded in archaeal genomes and also in T. maritima, A. aeolicus, and D. radiodurans.

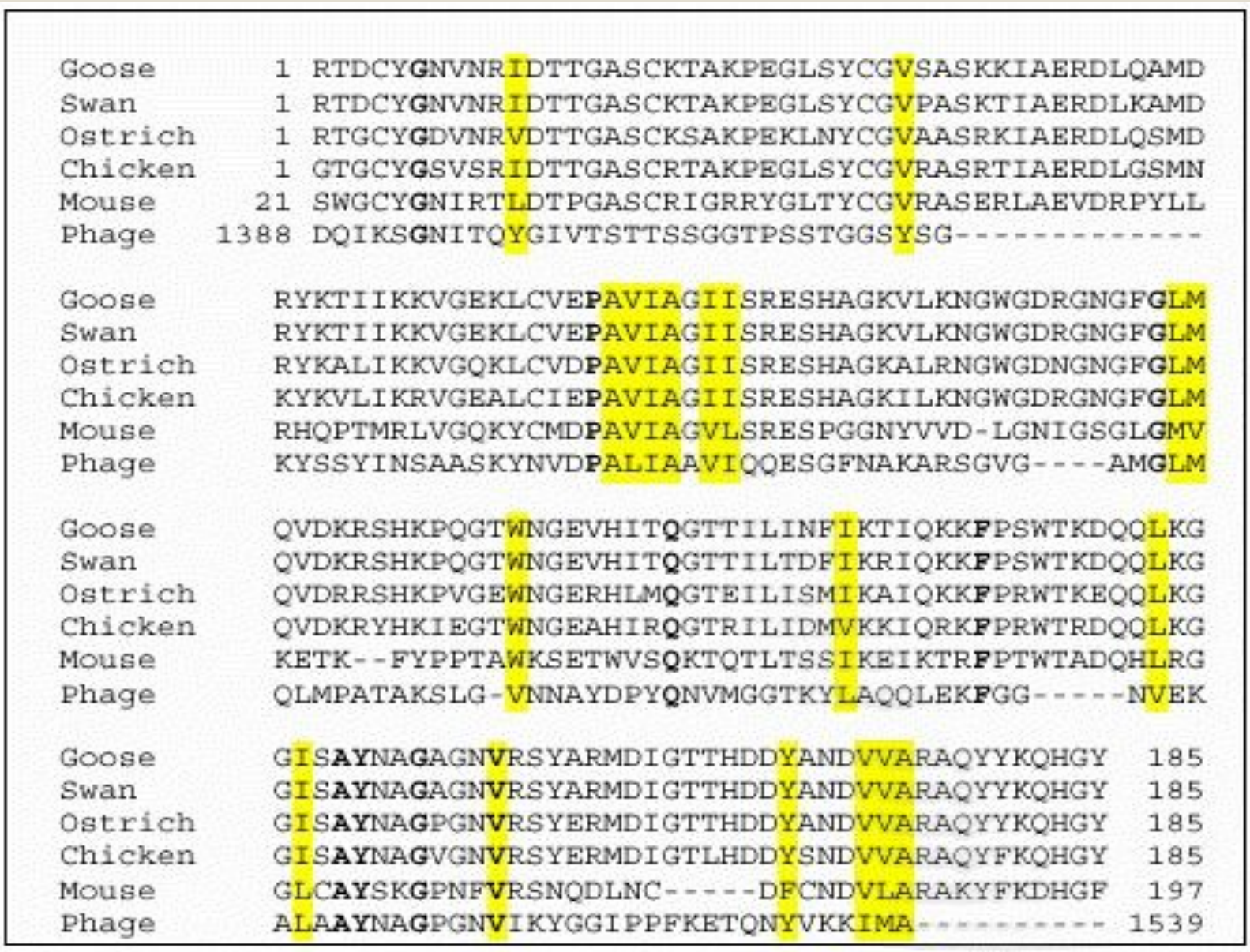


Fig : Movement of Glowworm to their Local Optima

References

1. K.N. Krishnanand, D. Ghose, "Detection of Multiple Source Locations using a Glowworm Metaphor with Applications to Collective Robotics", Swarm Intelligence Symposium, 2005, pp. 84-91.

2. Kaipa, K. and Ghose, D. (2017). *Glowworm Swarm Optimization Theory, Algorithms, and Applications.* Cham: Springer.

3. K.N. Krishnanand and D. Ghose. "Kinbots: A mobile robot platform for collective robotics applications," Technical Report GCDSL/2004/07, Department of Aerospace Engg., IISc, Bangalore, August 2004.

4. E. Bonabeau, M. Dorigo, G. Theraulaz. Swarm Intelligence: From Natural to Artificial Systems, Oxford University Press, 1999, pp.183-203.

5. Asha Gowda Karegowda and Mithilesh Prasad "A Survey of Applications of Glowworm Swarm Optimization Algorithm" International Conference on Computing and information Technology (IC2IT-2013)