*NBU Journal of*
**Plant Sciences**

Vol. 6, No. 1March 2012 p.9-17
ISSN No. 0974 6927

# Bioinformatics of pathogenic food bacteria

Uttam K Mondal[1], Arnab Sen[2] and Asim K Bothra[1]*

[1]Cheminformatics Bioinformatics Laboratory, Department of Chemistry, Raiganj College (University College), Raiganj-733134, Uttar Dinajpur, West Bengal, India. [2]NBU Bioinformatics Facility, Department of Botany, University of North Bengal, Siliguri-734013, West Bengal, India.

**Abstract**

Cholera is a fatal water borne disease caused by *Vibrio cholerae*. The primary symptoms include profuse watery diarrhea, muscle cramps and vomiting of clear fluid leading to rapid dehydration and electrolyte imbalance, and death in some cases. The causal organism i.e. *Vibrio cholerae* was first isolated by an Italian anatomist, Filippo Pacini in 1854. With the advent of genome sequencing technology, several *Vibrio cholerae* strains have been sequenced and available in public domain. This provided the opportunity for the scientists to study their genome in details. In the present review we have looked into the genetics of food pathogen as a whole, with special reference to research on *V. cholerae*. We have also critically reviewed various aspects of Bioinformatics like codon usage patterns, phylogenetic studies, studies on structural bioinformatics etc. on food borne bacteria. We are confident that the present review will help the researchers of food pathogenic biology to obtain valuable information about the present status of bioinformatics of food pathogenic bacteria and future outlook of these aspects.

**Keywords**: *Vibrio cholerae* , food pathogenic bacteria, codon usage, functional genomics

Cholera is a fatal disease caused by infection with the bacteria *Vibrio cholerae* (Ryan 2004) which is transmitted through contaminated water. Prior to the discovery of an infectious cause, the symptoms of cholera were thought to be caused by an excess accumulation of bile in the patient. The disease Cholera gets its name from the Greek word *choler* meaning bile. This was consistent with medical thought at the time, which held that four liquids or *humors* controlled health, and lead to such medical practices as bloodletting as a method of curing illnesses. The bacterium was first reported in 1849 by M. Gabriel Pouchet, who discovered it in stools from patients with cholera. But he did not appreciate its significance (William 1979). The first scientist to understand the significance of *Vibrio cholerae* was an Italian anatomist Filippo Pacini, who published detailed drawings of the organism in "Microscopical observations and pathological deductions on cholera" in 1854. He went on to publish additional papers in 1866, 1871, 1876 and 1880, all of which were ignored by the scientific community. He correctly described how the bacteria caused diarrhea, and developed treatments that were found to be effective (Bentivoglio & Pacini 1995). But his findings did not influence medical opinion. In 1874, scientific representatives from 21 countries voted unanimously to resolve that cholera was caused by environmental toxins from *miasmatas*, or clouds of unhealthy substances which float in the air (Howard-Jones N 1984).

Reports of pathogenic E. coli appeared in medical literature way back in 1947 (Ruchman & Dodd 1947). Publications regarding variants of E. coli which cause disease appeared regularly in medical journals throughout the 1950s, 60s, and 70s, (Macqueen 1954; Mc-Clure 1955; Gronroos 1957; Cowart & Thomason 1965; Linde *et al.*, 1966) with fatalities being reported in humans and infants starting in the 1970s (Glantz 1970; Drucker *et al.*, 1970; Smith & Gyles 1970). Infection with the bacteria *Helicobacter pyoli* is the cause of most stomach ulcers. The discovery is generally credited to Australian gastroenterologists Dr. Barry Marshall and Dr. J Robin Warren, who published their findings in 1983. The pair received the Nobel Prize in 2005 for their work. Before this, nobody really knew what caused stomach ulcers, though a popular belief was that the "stress" played a role. Some researchers suggested that ulcer was a psychosomatic illness (Paulley 1975; Kellner 1975; Aitken & Cay 1975). Besides, a number of food borne pathogens causing bacteria like *Salmonella* (non-typhoidal), *Listeria monocytogenes*, *Campylobacter*, (Mead *et al.*, 1999) were discovered. Genomic islands may contain large blocks of virulence determinants (adhesins, invasins, toxins, protein secretion systems, antibiotic resistance mechanisms, etc) and thus are referred to as pathogenicity islands. Pathogenicity islands were first described in pathogenic species of *E coli*, but since then have been found in the genomes of numerous bacterial pathogens of humans, animals, and plants (*Salmonella*, *Vibrio*, *Shigella*, *Yersinia*, *Listeria*, *S aureus*, etc) (Garcia *et al.*, 1999; Lindsay *et al.*, 1998). Pathogenicity islands are believed to have been acquired as a block by horizontal gene transfer because of their G+C content is significantly different from that of the genomes of the host micro-organism and they are often flanked by direct repeats. Many scientists are currently beginning to gain understanding of the molecular mechanisms involved in

*Corresponding author.
E-mail: asimbothra@gmail.com

the action of toxins. The good news is that there are a limited number of major toxin families that display common structural and biochemical motifs which may be exploited for future therapeutic development and these may be effective against multiple organisms.

**Genetics of food pathogen**

In 1906 the term "genetics" is introduced. Various techniques like mutations, deletion mapping, cloning vectors etc. have facilitated the identification of genes associated with food pathogens. The first genome of a food-borne bacterium to be sequenced was that of *C. jejuni* and it led to the discovery of important new aspects pertaining to the biology of this organism (Parkhill *et al.*, 2000). PAIs were first discovered in pathogenic *E. coli*. Pathogenicity islands of various pathogens are cag (*H. pylori*), SPI1-5 (*Salmonella* spp) VPI (*V. cholerae*) etc are responsible for different diseases (Hentschel & Hacker 2001). Kim *et al.*, designed a method for the Microarray detection of food-borne pathogens using specific probes prepared by comparative genomics (Kim *et al.*, 2008).

**Research trends in food pathogenic bacteria (FPB)**

The discovery of pathogenic organisms throughout the globe, studies concerning their gene products as well as the development of molecular biology helped FPB research enter a new stage where basic research blended well with latest sophisticated techniques. Techniques such as immune magnetic separation (IMS) and polymerase chain reaction (PCR) have paved the way for rapid and sensitive detection of food borne pathogens, and advances in nano-biotechnology have allowed for the miniaturization of devices. Collaborations between workers in the fields of engineering, nanotechnology and food science have introduced new lab-on-a-chip technologies permitting development of portable, hand-held biosensors for food pathogen detection. Despite the recent advancements in food pathogen detection, there still exist many challenges and opportunities to improve the current technology.

**Beginning of interdisciplinary research and dawn of bioinformatics**

It was in the year 1986 that the Department of Energy (DOE), USA and National Institutes of Health (NIH) started the Human Genome Project (HGP). It was one of the most happening experiments in the late 20$^{th}$ century. The objective of the project was to identify all of the genes in humans and craft a database containing the information (Ideker *et al.*, 2001). A number of other genome projects also started in major industrialized countries of Europe and Japan. The scientists were, at first, skeptical about the HGP, since huge amount of money had to be spent that would hamper basic research and as the project moved scientists worried about the massive amount of data and its interpretation (Bloom, 2001) In the year 2000, Prof. Collins of the National Human Genome Research Institute and Prof. Craig Venter of Celera Genomics appeared in a press conference and stated that they had achieved what was thought impossible and published the draft sequence of human genome (Wade 2000). That was the beginning and it opened the flood-gates for other genome sequencing projects. Gradually sequences of mouse, rat, worms and yeast were completed (Miller *et al.*, 2004). Kyrpides (1999) reported that in the end of the 20th century, there were 24 complete genomes that included 16 bacterial, 6 archaeal, and 2 eukaryotic genomes and currently there are more than two thousand genomes available in public databases (Kyrpides 1999). The large numbers of genomes resulted in the generation of huge amount of information concerning the genetic nature of biological organisms spanning different kingdoms, groups and lineages etc. Bloom proposed that the greatest problem appeared with the interpretation of underlying information from genomes leading to materialization of the new science of bioinformatics. Bioinformatics revolutionized the science of biology and directed it towards a more holistic approach compared to the reductionism visible in molecular biology research in the late 20$^{th}$ century (Bloom 2001). Now an organism is viewed as a system comprising of the information associated with genes and proteins that are responsible for maintaining day to day functions and networks of regulations that spell out how gene expression occur (Ideker *et al.*, 2001). In 1995, the publication of the complete genome of *Haemophilus influenzae* marked the beginning of another revolution in the field of bioinformatics (Fleischmann *et al.*, 1995). Currently there are about two thousand bacterial and archaeal genomes in the public domain. The publication of huge amount of sequence data helped in the development of high end computers, smart computing tools, for large-scale annotation, functional classification of the proteins (Searls 2000) and development of specific databases (Birney *et al.*, 2002) for availability to the broad scientific community. As the science of bioinformatics developed, computation became cheaper and cheaper and was duly complemented with the growth of Internet since the late 90s (Perez-Iratxeta *et al.*, 2006). High throughput tools were also developed by the beginning of the 21$^{st}$ century as genetic data became a gold mine for researchers (Perez-Iratxeta *et al.*, 2006). New software started developing for more efficient and comprehensive analysis of the genomes, proteomes and proteins. The stage was set and bioinformatics had already established itself as the leading science of the 21$^{st}$ century.

**Post genomic era and FPB**

The amalgamation of the knowledge of physiology, biochemistry, genetics and molecular biology gave idea about the understanding of the mechanism of host-pathogen interaction in pre-genomic era. The completion of the genomes of *Salmonella enterica* (McClelland *et al.*, 2001), *Helicobacter pylori Strain* G27 (Baltrus *et al.*, 2009) and sequences for a number of pathogenic bacteria. Pathogenicity islands were first described in pathogenic species of *E. coli*, but since then have been found in the genomes of numerous other bacterial pathogens of humans, animals, and plants (*Salmonella*, *Vibrio*, *Shigella*, *Yersinia*, *Listeria*, *S. aureus*, etc)

(Garcia *et al*., 1999; Lindsay *et al*., 1998). The first genome of a food-borne bacterium to be sequenced was that of *C. jejuni* and it led to the discovery of important new aspects (Parkhill *et al*., 2000). The studies on the genomes exposed new evidences pertaining to evolution and structure, interactions between host-pathogen interactions.

The research on functional genomics and proteomics with a focus on food pathogenic bacteria has been greatly significant in the 21st century. It has become necessary to investigate much on the comparative codon usage patterns, whole proteome analysis and molecular phylogeny using bioinformatics tools. Comparative genomics particularly focusing on codon usage using different parameters is expected to provide insight into the inherent molecular nature of the genomes of food pathogenic bacteria.

**Previous works on codon usage patterns**

In the post genomic era increasing number of genomes generated a concept among computational biologists that each and every genome has its own story. Since the time when the first nucleic acid sequences were obtained a number of hypothesis on the evolution of genomes have been put forward. The genetic code has been one of the most interesting aspects of biological science. The code is degenerate with multiple codons coding for a particular amino acid. Groups of codons coding for a particular amino acid are synonymous ones. It has been reported that these synonymous codons are somewhat conserved across species (Peden 1999). The increase in sequence information *albeit* partial in the 1980's facilitated the studies concerning the usage of synonymous codons in organisms. Majority of work on codon usage patterns at that period focused upon *E.coli* (Peden 1999). Gradually the techniques for codon usage were applied on mammalian, bacterial, bacteriophage, viral and mitochondrial genes (Grantham *et al*., 1980a; Grantham *et al*., 1981; Grantham *et al*., 1980b). On the basis of studies conducted on mRNAs from a number of prokaryotic and eukaryotic species, Grantham *et al*. (1980a) proposed the "Genome Hypothesis" which hypothesized that codon usage pattern of a particular genome was an explicit attribute of that organism. Grantham *et al*., (1981) reported that difference in codon usage pattern might be associated with the tRNA content. More work on the codon usage patterns (Gouy & Gautier, 1982) in *E.coli* regarding codon usage and tRNA abundance led to the conclusion that highly expressed genes exhibited non-random codon usage and used a small set of codons that corresponded to abundant tRNAs. However, it was not clear why specific synonymous codons were used preferably. Grosjean & Fiers (1982) opined that optimal codon choice is the outcome of the necessity imposed by interaction between codon and cognate tRNA. Ikemura (1981) defined optimal codon as one that was translated by the most abundant cognate tRNA which he later amended (Ikemura, 1985). These optimal codons are under the influence of translational efficiency. Kurland (1991) reported that translational efficiency is shaped by highest turnover of ribosomes, effectiveness of aminoacyl-tRNA and ternary complex conditions. Sharp *et al*. (1993) reported that the preference of some synonymous codons were the outcome of translational selection i.e., for increasing efficiency and accuracy a codon is used that is translated by the abundant tRNA species. Rocha (2004) correlated codon usage bias from the tRNA point of view. He proposed that co-evolution of tRNA gene composition and codon bias in genomes from tRNA's point of view concur with the selection-mutation-drift theory. A number of studies (Sharp *et al*., 1993; Carbone *et al*., 2005) revealed that codon bias is influenced by effective population size, translational selection, mutational pressure, compositional bias and genetic drift. Chen *et al*., (2004) postulated that codon bias is first and foremost influenced by mutational pressure and then translational selection. Studies on translational selection helped in the detection of highly expressed genes in genomes (Karlin & Mrazek, 2000). Majority of the highly expressed genes are associated with cell growth and cell division. It has been found that evaluating translational selection is quite difficult in genomes with a high or low G+C content owing to the consequence of intense levels of G+C content on codon usage. Medigue *et al*., (1991) applied the principles of codon usage, cluster analysis and correspondence analysis to study horizontal gene transfer mechanisms. A number of workers have suggested that codon usage and amino acid usage (Goldman & Yang 1994; Nesti *et al*. 1995; Pouwels & Leunissen 1994; Schmidt 1995) helps in determining the phylogenetic relationships between organisms. Besides codon usage, amino acid usage has also been studied extensively (Peden 1999). Ikemura (1981) was instrumental in reporting good correlation between amino acid composition and codon bias. Hydrophobicity, aromaticity and amino acid charges are testified to be influencial in amino acid usage (Lobry & Gautier, 1994). The prediction of open reading frames has been performed utilizing the information of codon usage (Krogh *et al*. 1994; Borodovsky *et al*. 1995). Peden (1999) provided some classic examples of the utilization of GeneMark prediction programme (Borodovsky *et al*., 1994) for identification of coding sequences from shotgun genome sequencing projects.

Sharp and Li (1987) were the pioneers in developing the Codon Adaptation Index (CAI) to assess the similarity among the synonymous codon usage of a gene to that of the reference set. This particular index has been commonly used as a parameter for predicting the highly expressed genes in an organism. Besides CAI, a number of indices are regularly used to investigate codon usage patterns in a number of organisms. Peden (1999) opined that two types of indices one evaluating on the whole deviation of codon usage from the expected set and the other determining bias towards a specific subset of optimal codons are used. Indices such as GC content, GC3 content (Peden 1999) effective number of codons (Nc) (Wright 1990), relative synonymous codon usage (RSCU) (Sharp *et al*., 1986) scaled chi-square and G statistic, P2 index (Gouy & Gautier, 1982) measuring the codon-anticodon interaction, codon bias index (CBI) (Bennetzen & Hall, 1982), Fop (frequency of optimal

codons) (Ikemura, 1985) and codon adaptation index are very significant in studies concerning codon usage patterns. Besides these indices, correspondence analysis (Benzecri, 1992) a type of multivariate statistical analysis has also been commonly used to find out the degree of associations between different genes and amino acids. It is a technique that explores non random synonymous codon usage.

This important field of study starved due to the scarcity of ophisticated software in the 1980's. The basic problem was the integration of different indices into one program that would serve the researchers well. The foremost software was CODONS (Lloyd & Sharp, 1992). The GCG package (GCG, 1994) consisted of programs for analysis codon usage. CORRESPOND was used for performing multivariate statistical analysis (Peden 1999). NetMul (Thioulouse *et al.*, 1995) was designed as a subset of ADE software for studying of codon usage and multivariate statistical analysis. CODON W (Peden 1999) took over the stage and became very popular because of its robust nature and error free analysis. INCA (Supek & Vlahovicek, 2004) is also being used by researchers. Software was also developed for codon usage optimization. This involved alteration of rare codons in target genes so as to imitate the codon usage of host with no modification of the amino acid sequence of encoded proteins (Gustafsson *et al.*, 2004). GeneDesign (Richardson *et al.*, 2006), Synthetic Gene Designer (Wu *et al.*, 2006), Gene Designer (Villalobos *et al.*, 2006) are some of the useful software packages providing a platform for synthetic gene design and codon optimization. There has been a number of softwares for estimation of the highly expressed genes employing CAI. CODON W (Peden 1999), JCAT (Grote *et al.*, 2005), CAI CALCULATOR (Wu *et al.*, 2005), ACUA (Umashankar *et al.*, 2007) and e CAI server (Puigbo *et al.*, 2008) are commonly used. However, the e CAI server (Puigbo *et al.*, 2008) has proved to be the most powerful and efficient tool for estimation of expression levels of the genes.

Although a lot of work has been performed and is going on in codon usage of different microorganisms but very little work has been performed on codon usage of food pathogenic bacteria. The first species in which codon usage was examined in detail, the bacterium Escherichia coli (Post & Nomura 1980; Ikemura 1981) and the yeast Saccharomyces cerevisiae (Ikemura 1982; Bennetzen & Hall 1982), were both found to show strong evidence of natural selection on codon usage. All these are just very little and a comprehensive work with focus on lifestyle patterns is the need of the hour.

**Phylogenetic studies on FPB**

Nothing in biology makes sense without evolution. This also applies for food pathogenic bacteria too. To establish the differentially evolved genes Eswarappa *et al.*, compared the phylogeny of the nine potential differentially evolved genes with phylogeny of the S. eneterica species and with the phylogeny of the five pathogenecity island of Salmosella (SPI-1 to SPI-5) (Eswarappa *et al.*, 2008). This also applies for other pathogenic bacteria.

**Structural bioinformatics of FPB**

High throughput sequencing of a number of eukaryotes and a number of bacteria has paved way to the field of structural bioinformatics. Structural bioinformatics is expected to give rise to a large number of protein structures. The technological benefits of this field are already yielding results and have a substantial impact on structural biology research for prokaryotes (Burley & Bonnano, 2002).Metropolis *et al.*, introduced the Monte-Carlo technique to the solving of physical equations. It described the idea of using random numbers to project a representative subset of conformational space, whilst using the exponent of the energy as a probability filter (Metropolis *et al.*, 1953). Levitt and Warshel simulated the folding of the Bovine Pancreatic Trypsin Inhibtor (Levitt & Warshel 1975). Case and Karplus work on "Dynamics of ligand binding to heme protein" in 1979. This is arguably the first simulation of ligand moving through the protein. At the early stage, either the protein could be made stationary allowing the oxygen to bounce around, or let individual sidechains could be allowed to hit by the oxygen. Each oxygen atom was simulated for 3.75 ps. For these pioneers it was a surprise to see that the oxygen bounces around the inside of the myoglobin, without getting too far. Nevetheless, they identified 2 different pathways for the oxygen to travel into the binding site (Case & Karplus 1979). Umbrella sampling is the most popular method of exploring large conformational changes in MD. In 1998 Notthrup et al show "Dynamical theory of activated processes in globular proteins". In this paper, Karplus and friends modeled a rather more modest conformational change: the swinging of an aromatic residue sidechain. From the simulations, they generated a free-energy surface, from which they calculated a sidechain flipping rate. This paper is important not just for simulating the first sidechain flip but also for introducing proteins to "umbrella sampling" (Northrup *et al.*,1982). First application of normal modes to identify low-frequency oscillations using the energy minimum of the molecular mechanics force-field of a protein was deseribed by Brooks & Karplus in 1983 (Brooks & Karplus *et. al.*, 1983). This is the basic technique to identify domain-level motions in a protein. First simulation of a protein in explicit waters was done by Levitt & Sharon (Levitt & Sharon 1988). Suddenly, acceptable computer resources got a whole lot more expensive. Guilbert *et. al.* gave first description of the RMSD potential, a powerful method for identifying low-energy pathways in the neighborhood of a given static structure (Guilbert *et al.*, 1995). Israelachvili & Wennerström showed that water molecules can have structuring effects of several Ångstroms flagging the importance of using explicit water molecules (Israelachvili & Wennerström 1996). The first reported 1 microsecond MD simulation, it was a mamoth effort for the late 90's, really pushing the technology of parallel clusters, a technology which the whole scientific community accepts. They tried to fold a tiny protein, the villin headpiece subdoamin, and got some of the way (Duan & Kollman 1998). Bernèche and

Roux uses umbrella sampling to identify all the positions of the K+ ion along the KcsA K+ membrane channel. They show that the channel for K+ is virtually barrierless, hence it is a diffusion controlled process but more impressively, they identify two K+ sites just outside the channel, which was subsequently identified by electron density in a high-resolution structure (Bernèche & Roux 2001).

Rapid developments of technological know how in proteomics coupled with the improvement of bioinformatics tools have resulted in a deluge of structural information that guarantees acceleration in research. *In silico* identification of potential therapeutic targets in *Clostridium botulinum* by the approach of subtractive genomics was carried out by Koteswara *et al.*, (2010). Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. It has been shown that three-dimensional protein structure is evolutionarily more conserved than expected due to sequence conservation (Kaczanowski 2010).

Proteins fulfill several crucial functions, having catalytic, structural and regulatory roles in all organisms. Knowledge of the three dimentional structure of proteins is a basic prerequisite for understanding their function. It provides a basis for studies of substrate or ligand interactions with a particular enzyme or protein. Three dimensional structures of proteins are determined by X-ray crystallography and NMR spectroscopy. By 1st June 2007, more than 40000 sets of atomic coordinates for proteins have been deposited in the Protein Databank (Berman 2008).

The first modeling studies, carried out in the late 1960s and early 1970s, were based upon the construction of wire or plastic models. The later studies were performed using interactive computer graphics. Browne *et al.* published the first report on homology modeling (Browne *et al.*, 1969). The model was bovine α-lactalbumine on the known 3D structure of hen egg white lysozyme. Later on Warme *et al.* produced a model for α-lactalbumine on the basis of the crystal structure of lysozyme (Warme *et al.*, 1974). These models were constructed by taking the existing coordinates of the known structure, and mutating side chains not identical in the protein to be modeled. This approach to protein modeling is still employed today with considerable success, especially when the proteins are similar (May & Blundell 1994).

McLachlan and Shotton modeled α-lytic proteinase of the fungus Myxobacter 495 on the basis of the structures of the mammalian chymotrypsin and elastase (McLachlan & Shotton 1971). This was a difficult task because the sequence identity between the protein to be modeled and the known structures was of the order of 18%. Subsequently, the crystal structure of α-lytic proteinase was determined and comparison was made between the X-ray structure with the homology model. It was found found that although segment of both domains of the model were built correctly, misalignment of the sequence led to local errors.

Greer introduced the modeling of variable regions in proteins on the basis of equivalent region from homologous proteins of known structures. In order to construct the homology models of a number of different serine proteases, he superimposed the structures of trypsin, chymotrypsin and elastase and found many equivalent Cα atoms within 1.0 A of one another (Greer 1980, 1981). The regions comprised of the amino acids of these Ca atoms were described as structurally conserved regions (SCRs). All of the remaining positions corresponded to structurally variable or loop regions (VR) where the insertions/deletions were located. The main chain of both structurally conserved (SCR) and variable regions (VR) were built from the fragments of known serine proteases. The side chains were modeled according to the conformation found at the equivalent positions for those identical side chains in the known structures.

Among the aspartic proteinases, initially models were constructed for rennin and rennin inhibitor complexes using the 3D structure of the distantly related fungal proteinases (1985; Akahane *et al.*, 1985). Later on, the homology models for rennin were built using the structures of mammalian aspartic proteases, pepsin and chymosin (Fragao *et al.*, 1994; Hutchins and Greer 1991). Comparison of the rennin models constructed from fungal and mammalian enzyme revealed that errors in the models arose from the difference in the arrangement of helices and strands between the mammalian and fungal aspartic proteinases, as well as the different variable regions. Nevertheless, the active of rennin was modeled reasonably correctly.

In the early eighties, manual homology modeling was facilitated by manipulation of protein molecules on the graphics terminal that was made possible by computer programs such as FRODO (Jones 1978). Since mid 1980s, a large number of homology models of proteins with different folds and functions have been reported in the literature (Johnson *et al.*, 1994; Sali 1995).

The sequence alignment and template structure were then used to produce a structural model of the target. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity (Marti-Renom *et al.*, 2000).

The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The approach can be complicated by the presence of alignment gaps (commonly called indels) that indicate a structural region present in the target but not in the template, and by structure gaps in the template that arise from poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure. Model quality declines with decreasing sequence identity; a typical model has ~1-2 Å root mean square deviation between the matched $C^{\alpha}$ atoms at 70% sequence identity but only 2-4 Å agreement at 25% sequence identity. However, the errors are significantly higher in the loop regions, where the amino acid sequences of the target and template proteins may be

completely different.

Homology modeling can produce high-quality structural models when the target and template being closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds (Williamson 2000). Like other methods of structure prediction, current practice in homology modeling is assessed on a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

**Future prospects**

Food pathogenic bacteria research presents a completely new outlook in the post genomic era. The wealth of information obtained from the genome projects needs to be mined. New insights are gained with the discovery of novel toxin genes. Bioinformatics, in combination with metagenomics as well as metaproteomics based approaches, has the potential to give a more detailed scenario that underlies the mechanisms of pathogenicity as well as diseases. In absence of crysllographic or NMR structure Homology modeling will enlighten us about three dimensional structure and molecular dynamics simulation providing newer possibilities for exploring the molecular mechanism and activity of that toxin. Thus scientific perception will continue to play a vital role in creating models that clarify the functions of pathogens in improved manner.

**References**


Aitken C & Cay E (1975). Clinical psychosomatic research. *International journal of psychiatry in medicine*, 6: 29–41.

Akahane K Nakagava S & Umeyana H Moriguchi I, Hirose S, Iizuka K & Murakami K (1985). Three-dimensional structure of human rennin. *Hypertension*, 7:3-12.

Baltrus DA, Amieva MR, Covacci A, Todd M. Lowe TM, Merrell DS, Ottemann KM, Stein M, Salama NR & Guillemin K (2009). The Complete Genome Sequence of *Helicobacter pylori* Strain G27. *Journal of Bacteriology*, 191:447-448.

Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Cryst. (2008). A64, 88–95.

Benedito VA, Dai X, He J, Zhao XP & Udvardi MK (2006). Functional genomics of plant transporters in legume nodules. *Func Plant Biol* 33:731-736.

Bennetzen JL & Hall BD (1982). Codon selection in yeast. *J. Biol. Chem.*, 257:3026–3031.

Bentivoglio M & Pacini P (1995). Filippo Pacini: a determined observer. *Brain Res. Bull.* 38: 161–165.

Benzecri JP (1992). Correspondence analysis handbook. Marcel Dekker, New York.

Bernèche S & Roux B (2001). Energetics of ion conduction through the $K^+$ channel. *Nature*, 414:73-77.

Birney E, Clamp N & Hubbard T (2002). Databases and tools for browsing genomes. *Annu Rev Genomics Hum Genet*, 3:293-310.

Bloom M (2001). Biology in silico: The Bioinformatics revolution. *The American Biology Teacher*, 63: 397-403.

Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Medigue C & Danchin A (1995). Detection of new genes in a bacterial genome using Markov models for 3 gene classes. *Nucleic Acids Res*, 23: 3554-3562.

Borodovsky M, Rudd KE & Kounin EV (1994). Intrinsic and Extrinsic Approaches for Detecting Genes in a Bacterial Genome. *Nucleic Acids Res*, 22: 4756-4767.

Brooks B & Karplus M (1983). Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *PNAS*, 80:6571-6575

Browne WJ, North ACT, Philips DC, Brew K, Vanaman TC & Hill RC (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme, J. *Mol. Biol.*, 42: 65-86.

Burley, SK & Bonnano, JB (2002). Structuring the universe of proteins. *Annu Rev Genomics Hum Genet*. 3:243-262.

Carbone A, Kepes F & Zinovyev A (2005). Codon bias signatures, organization of microorganisms in codon space and lifestyle. *Mol Biol Evol*. 22: 547-561

Case DA & Karplus M (1979). Dynamics of ligand binding to heme protein. *J Mol Biol*. 132:343-368.

Chen SL, Lee W, Hottes AK, Shapiro L & McAdams HH (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*, 101: 3480-3485.

Cowart GS & Thomason BM (1965). Immunofluorescent detection of Escherichia coli; incidence of certain serogroups suspected of being pathogenic. *Am. J. Dis. Child*. 110: 131–136.

Drucker MM, Polliack A, Yeivin R & Sacks TG (1970). Immunofluorescent demonstration of enteropathogenic Escherichia coli in tissues of infants dying with enteritis. *Pediatrics*. 46: 855–864.

Duan Y& Kollman PA (1998). Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science*, 282:740-744.

Eswarappa SM, Janice J, Nagarajan AG, Balasundaram SV, Karnam G, Dixit NM & Chakravortty D (2008). Differentially Evolved Genes of *Salmonella* Pathogenicity Islands: Insights into the Mechanism of Host Specificity in *Salmonella* *PLoS ONE*.3: e3829 doi: 10.1371/ journal.pone.0003829.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF Kerlavage AR, Bult CJ, Kerlavage AR, Tomb JF, Dougherty BA, Merrick JM & *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269:496–512.

Fragao C, Topham C, Dhanaraj V & Blundell TL (1994). Comparative modelling of human renin: A retrospective evaluation of the model with respect to the X-ray crystal structure. *Pure Appl. Chem.*, 66:43-50.

Garcia E, Llull D, Lopez R (1999). Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. *Int Microbiol.*, 2:169–76.

GCG (1994). Programme Manual for the Wisconsin Package, Version 8, University of Wisconsin.

Glantz PJ (1970). Unclassified Escherichia coli serogroup OX1 isolated from fatal diarrhea of rabbits. *Can. J. Comp. Med.*, 34: 47–49.

Goldman N & Yang ZH (1994). Codon based model of nucleotide substitution for protein coding DNA sequences. *Mol Biol Evol*,11: 725-736.

Gouy M & Gautier C (1982). Codon usage in bacteria correlation with gene expressivity. *Nucleic Acids Res.* 10: 7055-7074.

Grantham R, Gautier C & Gouy M (1980a). Codon frequencies in 119 genes confirm consistent choices of degenerate base according to genome type. *Nucleic Acids Res.*, 8:892-1912.

Grantham R, Gautier C, Gouy M, Jacobzone M & Mercier R (1981). Codon catalogue usage is a genome strategy for genome expressivity. *Nucleic Acids Res.*, 9: r43-r75.

Grantham R, Gautier C, Gouy M, Mercier R & Pave A (1980b). Codon catalogue usage and the genome hypothesis. *Nucleic Acids Res.* 8: r49-r62.

Greer J (1980). Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sc. USA*, 77:3393-3397.

Greer J (1981) Comparative Model-building of the Mammalian Serine Proteases. *J. Mol.Biol.*, 153:1027-1042.

Gronroos JA (1957) Investigations on Escherichia coli O groups 1-25, 44 and 78 and serotypes 26:B6, 55:B5, 86:B7, 111:B4, 125:B15 and 126:B16; occurrence in faeces of healthy and diarrhoeal infants. *Annales medicinae experimentalis et biologiae Fenniae*, 35: 1–35.

Grosjean H & Fiers W (1982). Preferential codon usage in prokaryotic genes-the optimal codon anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18: 199-209.

Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC & Jahn D (2005). JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res*, 33:W526-31.

Guilbert C, Perahia D & Mouawad L (1995). A method to explore transition paths in macromolecules. Applications to hemoglobin and phosphoglycerate kinase. *Computer Physics Communications*, 91:263-273 .

Gustafsson C, Govindarajan S & Minshull J (2004). Codon bias and heterologous protein expression. *Trends Biotechnol.* 22:346-353.

Hentschel U& Hacker J (2001). Pathogenicity islands: the tip of the iceberg. *Microbes and Infection*, 3:545–548.

Howard-Jones N (1984). Robert Koch and the cholera vibrio: a centenary. *British medical journal (Clinical research ed.)*, 288: 379–381.

Hutchins C & Greer J (1991). Comparative modeling of proteins in the design of novel renin inhibitors. *Crit. Rev. Biochem. Mol. Biol.*, 26:77-127.

Ideker, T, Galitski, T & Hood, L (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2: 343-372.

Ikemura T (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons of its protein genes. *J Mol Biol*, 146:1-21.

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *J. Mol. Biol.*, 158:573–597.

Ikemura T (1985). Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 2:13-34.

Israelachvili J, & Wennerström *H* (1996). Role of hydration and water structure in biological and colloidal interactions. *Nature*, 379:21-225.

Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL (1994). Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.*, 29: 1-68.

Jones TA (1978) A graphics model building and refinement system for macromolecules. *J. Appl.* Cryst. 11: 268-272.

Kaczanowski S & Zielenkiewicz P (2010). Why similar protein sequences encode similar threedimensional structures? *Theoretical Chemistry Accounts* 125:543-50.

Karlin S & Mrazek J (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* 182:5238-5250.

Kellner R (1975) Psychotherapy in psychosomatic disorders. *Arch. Gen. Psychiatry*, 32: 1021–1028.

Kim HJ, Park SH, Lee TH, Nahm BH, Kim YR, Kim HY (2008). Microarray detection of food-borne pathogens using specific probes prepared by comparative genomics. *Biosens Bioelectron.*, 24 (2):238-246.

Koteswara Reddy G, Nagamalleswara Rao K, Phani Rama Krishna B & Aravind S (2010). *In silico* identification of potential therapeutic targets in *Clostridium botulinum* by the approach subtractive genomicsInternational. *Journal of Bioinformatics Research*, 2:12-16.

Krogh A, Mian IS & Haussler D (1994). A hidden Markov model that finds genes in *Escherichia coli* DNA. *Nucleic Acids Res.* 22: 4768-4778.

Kurland CG (1991). Codon bias and gene-expression. *FEBS Letters*, 285:165-169.

Kyrpides NC (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide. *Bioinformatics* 15:773–774.

Levitt M & Warshel A (1975). Computer simulation of protein folding. *Nature*. 253:694-698.

Linde K, Köditz H & Hasenjäger H (1966). Treatment of infantile diarrhea caused by pathogenic Escherichia coli strains with oral large doses of benzyl procaine penicillin (in Polish). *Pediatria polska*, 41: 905–12.

Lindsay JA, Ruzin A, Ross HF, Kurepina N & Novick RP (1998). The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus. *Mol Microbiol.*, 29:527–543.

Lloyd AT & Sharp PM (1992). CODONS - a microcomputer program for codon usage analysis. *J Heredity*, 83: 239-240.

Lobry JR & Gautier C (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. *Nucleic Acids Res.* 22: 3174-3180

M Levitt M & Sharon R (1988). Accurate simulation of protein dynamics in solution. *PNAS*, 85:7557-7561.

Macqueen RL (1954). Isolation of Bact. coli O:26.B.6 from a child with recurrent diarrhea. *British medical journal*. 1:1475–1476.

Martin-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F &

Sali A (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325.

May ACW & Blundell TL (1994). Automated comparative modelling of protein structures. *Curr. Opin. Biotechnol*, 5:355-360.

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P,Courtney L, Porwollik S, Ali J, Dante M, Du F et al., (2001). Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, 413:852–856.

McClure WB (1955). A severe nursery epidemic of diarrhoea associated with Esch. coli type 111 B4. *Canadian Medical Association journal*, 72:83–88.

McLachlan AD & Shotton DM (1971). Structural similarities between α-lytic protease of myxobacter 495 and elastase. *Nature*, 229:202-205.

Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM & Tauxe RV (1999). Food-related illness and death in the United States. *Emerg Infect Dis.*, 5:607-625.

Medigue C, Rouxel T, Vigier P, Henaut A & Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol*, 222: 851-856.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087-1092.

Miller W, Kateryna DM, Nekrutenko A & Hardison RC (2004). Comparative genomics. *Annu Rev Genomics Hum Genet*, 5:15-56.

Nesti C, Poli G, Chicca M, Ambrosino P, Scapoli C Brrai I (1995). Phylogeny inferred from codon usage pattern in 31 organisms. *Comp Appli Biosciences*, 11:167-171.

Northrup SH, Pear MR, Lee CY, McCammon JA & Karplus M (1982). Dynamical theory of activated processes in globular proteins. *PNAS*, 82:4035-4039.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MTG, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S & Barrell BG (2001). Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature*, 413:848–852.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T & Holroyd S et al (2000). The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature*, 403:665–668.

Paulley JW (1975) Cultural influences on the incidence and pattern of disease. *Psychotherapy and psychosomatics*, 26: 2–11.

Paden J (1999). Analysis of codon usage. *PhD thesis*, University of Nottingham.

Perez-Iratxeta, C, Andrade-Navarro, MA & Wren, JD (2006). Evolving research trends in bioinformatics. *Brief In Bioinform* 8: 88-95

Post LE & Nomura M (1980). DNA sequences from the str operon of Escherichia coli. *J. Biol. Chem.*, 255:4660–4666.

Pouwels PH & Leunissen JAM (1994). Divergence in codon usage of *Lactobacillus* species. *Nucleic Acids Res*, 22: 929-936.

Puigbo P, Bravo IG & Vallve-Garcia S (2008). E-CAI: a novel server to estimate an expected value of codon adaptation index (e CAI). *BMC Bioinformatics*, 9: 65 doi 10.1186/1471-2105-9-65.

Richardson SM, Wheelan SJ, Yarrington RM & Boeke JD (2006). GeneDesign: Rapid, automated design of multi kilobase synthetic genes. *Genome Res*, 16: 550-556.

Rocha EP (2004). Codon usage bias from tRNA's point of view: redundancy, specialization and efficient decoding for translational optimization. *Genome Res*, 14: 2279-2286.

Ruchman I & Dodd K (1947). The Isolation of a Strain of Escherichia coli Pathogenic for the Rabbit's Eye from a Patient with Diarrhea. *J. Bacteriol.*, 53:653–656.

Ryan KJ (2004). *Sherris Medical Microbiology, 4th ed. McGraw Hill.*

Sali A (1995). Modelling mutations and homologous proteins. *Curr. Opin. Biotechnol.*, 6: 437-451.

Schmidt W (1995) Phylogeny reconstruction for protein sequences based on amino acid properties. *J Mol Evol.*, 41: 522-530.

Searls DB (2000). Bioinformatics tools for whole genomes. *Annu Rev Genomics Hum Genet*, 1:251-279.

Sharp PM, Stenico M, Peden JF & Lloyd AT (1993). Codon usage: Mutational bias, translational selection, or both? *Biochem Soc Trans*, 21:835- 841

Sharp PM, Tuohy TMF & Mosurski KR (1986). Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*, 14: 5125-5143.

Sharp, PM & Li WH (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15:1281-1295.

Smith HW & Gyles CL (1970). The relationship between two apparently different enterotoxins produced by enteropathogenic strains of Escherichia coli of porcine origin. *J. Med. Microbiol.*, 3: 387–401.

Supek F & Vlahovicek K (2004). INCA: Synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, 20: 2329-2330

Thioulouse J, Doledec S, Chessel D & Oliver JM (1995). ADE softeware: multivariate analysis and graphical display of environmental data. In *Sofware per l'Ambiente*, edited by Guariso G & Rizzoli A Patron editor, Bolonia.

Umashankar V, Arun Kumar V & Sudarsanam D (2007). ACUA: A software tool for automated codon usage analysis. *Bioinformation*, 2: 62-63.

Villalobos A, Ness JE, Gustafsson C, Minshull J & Govindarajan S (2006). Gene designer: A synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics*, 7: 285

Wade N (2000). Genetic code of human life is cracked. *New York Times*, June 27.

Warme PK, Momany FA, Rumball SU, Tuttle RW & Sherago

HA (1974). Computation of structures of homologous proteins. alpha-lactalbumin from lysozyme. *Biochemistry*, 13:768-782

William B (1979). *The History of Bacteriology, Dover publications.*

Williamson AR. (2000). Creating a structural genomics consortium. *Nat Struct Biol* , 7 S1 (11s):953.

Wright F (1990). The effective number of codons used in a gene. *Gene*, 87: 23-29

Wu G, Bashir-Bello N & Freeland SJ (2006). The synthetic gene designer: A flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr Purif*, 47:441-445

Wu G, Culley DE & Zhang W (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Micobiologym*, 151: 2175-2187