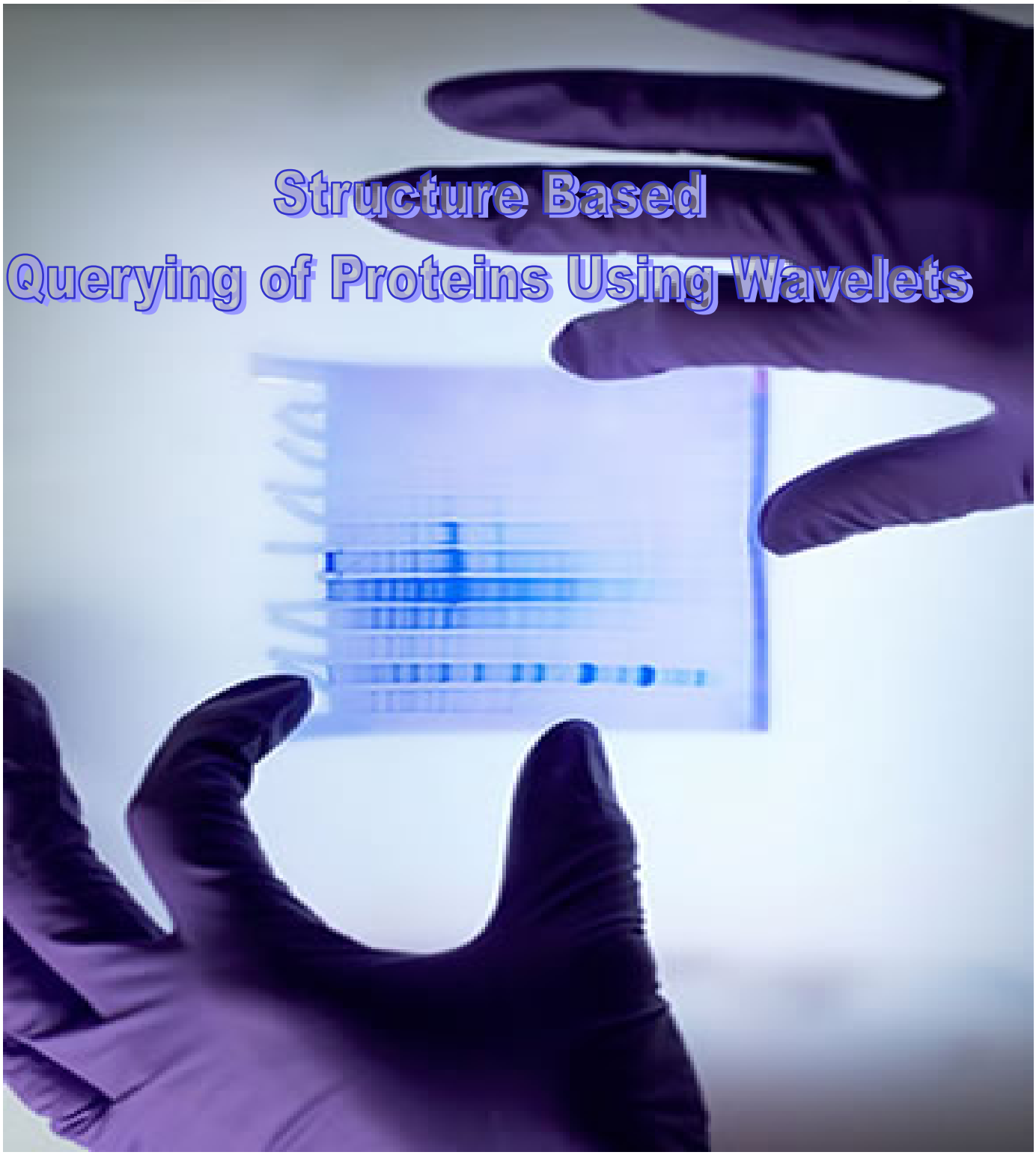


1

Supervisor

Dr.Eng: Nabil Lashin

TEAM MEMBERS

1 – Mohammed Mamdouh Mohammed .

2 – Elbasheer Mohammed Elshourbagy

3 – Ahmed El-Modather Mohammed .

4 – Adhem Kamel Mohammed .

5 – Evram Adly Moawad .

6 – Ahmed Mohammed Abdelmenem .

2

Acknowledgments

We are indebted to a number of individuals in academic circles as well as in faculty who have contributed to complete this project successfully.

Their contributions have *been important* in so many different ways that we find it difficult *to* acknowledge them in any other manner.

In general we wish to *extend* our *appreciation to our supervisor* *Dr.Nabil Lashin* *for* his suggestions and the valuable assistance during the *project* on how to improve our work and complete it more useful to others and who gives us the motivation along the right path. Also we acknowledge the efforts of *Eng.Hassan Abas* and **his continuous encouragements and advices**.

Finally, *We should thank ALLAH for charging us with the all faith and the power to make us people who are depended on. We can't forget thanking our fathers,help us to get our destinations.*

Thank you.....

Dr. Nabil Lashin.

3

Index

Ch1: Introduction

1.1 Objectives of the project

1 - As the number of proteins increased rapidly in the last few decades, it is too hard to search in the protein data bases according to the large data used which causes much time in searching and matching proteins to know the type of the proteins.

2 - We tried to use some computer algorithm to search in the databases much faster and to get the most closely types of the proteins.

3 - So we used the wavelet techniques and the Dataminig to search in the Protein Data Bank (PDB).

4 - This project is very useful for many importany applications that deals with the amino acids.

5 - This project can be used instesd of an expensive device which uses the beam of X-rays to get the protein types easily and much faster.

1.2 What is Protein?

A protein is a long train of amino acids linked together. Proteins have different functions; they can provide structure (ligaments, fingernails, hair), help in digestion (stomach enzymes), aid in movement (muscles), and play a part in our ability to see (the lens of our eyes is pure crytalline protein).

Protein is a long chain molecule made up of amino acids joined by peptide bonds. Protein forms the structural material of bodily tissues.

Proteins, the principal constituents of the protoplasm of all cells, are of high molecular weight and consist essentially of combinations of a amino acids in peptide linkages.

7

Twenty different amino acids are commonly found in proteins and each protein has a unique, genetically defined amino acid sequence which determines its specific shape and function.

They serve as enzymes, structural elements, hormones, immunoglobulins, etc. And are involved in oxygen transport, muscle contraction, electron transport and other activities throughout the body and in photosynthesis. (explained in Ch2).

1.3 What is Amino acid?

Amino acids are molecules that contain at least one amine group (-NH2) and at least one carboxylic acid group (-COOH). When these groups are both attached to the same carbon, the acid is an -amino acid. -amino acids are the basic building blocks of proteins.

1.4 What is Data mining?

Data mining: is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data and "the science of extracting useful information from large data sets or databases. Data mining in relation to enterprise resource planning is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making.

Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities. However, abdicating control of this process from the statistician to the machine may result in false-positives or no useful results at all. (explained in Ch4).

8

1.5 What is Wavelet?

A wavelet - is a kind of mathematical function used to divide a given function or continuous-time signal into different frequency components and study each component with a resolution that matches its scale.

The wavelet analysis procedure is to adopt a wavelet prototype function, called an analyzing wavelet or mother wavelet. Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients. And if you further choose the best wavelets adapted to your data, or truncate the coefficients below a threshold, your data is sparsely represented. This sparse coding makes wavelets an excellent tool in the field of data compression. (explained in Ch5).

1.6 Problems Faced during the project

1 - Less resouces of the faculty.

2 - Informatin was hard to find.

3 - We tried to coordinate with the Ministery of Medicine but they were not helpful.

4 - To gain the most useful use of this project, we were suppose to use the project on a huge real world data but we just used it in a small data during to the less of the resources.

1.7 Hard and Software used

1.7.1 Hardware

an ordinary Desktop Computer can deals with the project easily but ofcourse A faster one can search in the database rabidly and

9

do the wavelet techniques rapidly, so the results can be implemented much faster.

1.7.2 Software

1 - An ordinary Oberating System like WINDOWS XP.

2 - The Environment used to program the algorithm of the project is (.NET FrameWork - 2 -2005).

3 - The Programming Language Used was C#.

1.8 Skills needed

1 - The Great ability to deal with the compiler and the frame work easily.

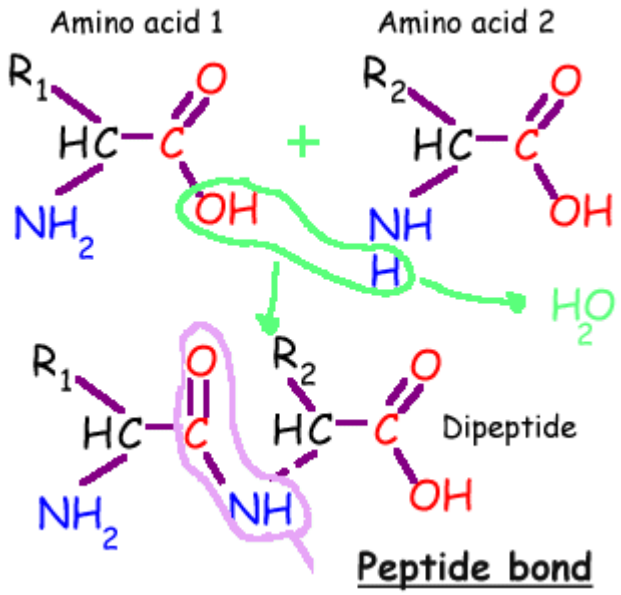
2 - The great ability to code the project using (C#) language and to ably the language in such a way that couldn't use much resource.

10

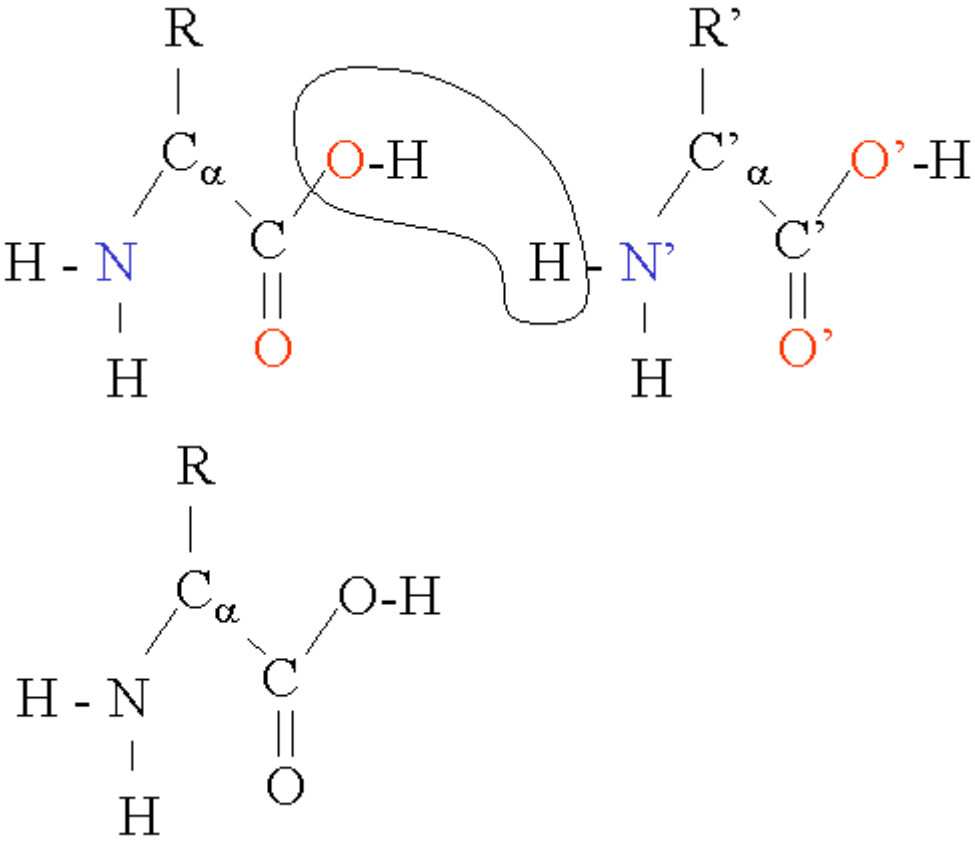
Ch 2: *PROTEIN* STRUCTURE

2.1 Structure of Amino Acids

An α-amino acid consists of a part that is present in all the amino acid types, and a side chain that is unique to each type of residue. The C_α atom is bound to 4 different molecules,n an amino group, a carboxy group, a hydrogen and a side chain, specific for this type of amino acid. An exception from this rule is proline, where the hydrogen atom is replaced by a bond to the side chain. Because the carbon atom is bound to four different groups it is chiral, however only one of the isomers occur in biological proteins. Glycine however, is not chiral since its side chain is a hydrogen atom. A simple mnemonic for correct L-form is "CORN": when the C_α atom is viewed with the H in front, the residues read "CO-R-N" in a clockwise direction



11



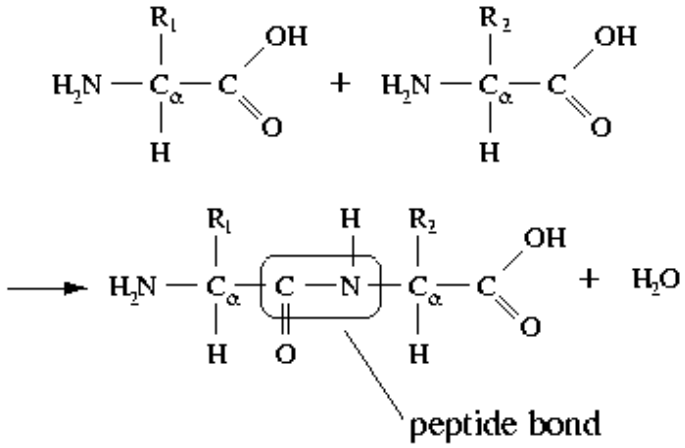
2.1.1 Co – R – N Role

The side chain determines the chemical properties of the α-amino acid and may be any one of the 20 different side chains:
The subunits of a protein are amino acids or to be precise **amino acid residues**. An amino acid consists of a central carbon atom (the alpha Carbon C_{alpha}) and an amino group (NH_2), a hydrogen atom (H), a carboxy group (COOH) and a side chain (R) which are bound to the C_{alpha}. Different **side chains** (R_i) make up different amino acids with different physic-chemical properties. Proteins are made out of 20 amino acids.

2.1.2 Peptide bond

is formed via covalent binding of the Carbon atom of the Carboxy group of one amino acid to the nitrogen atom of the amino group of another amino acid by dehydration:

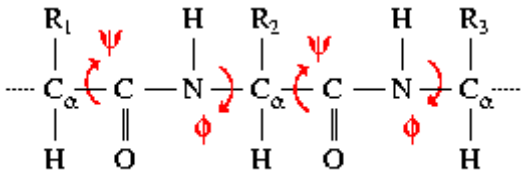
12

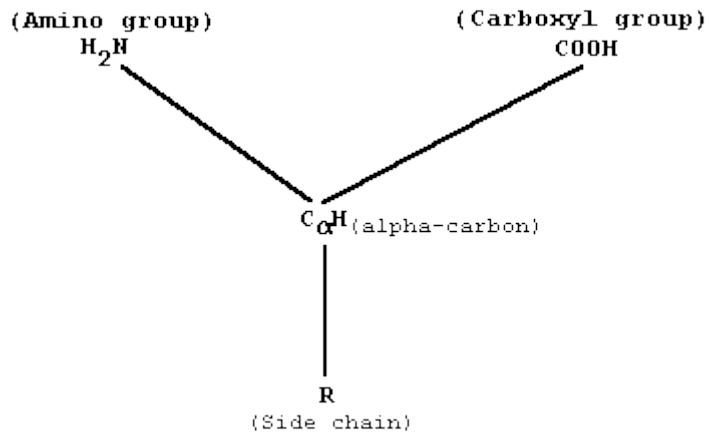


Peptide bond linking two amino acids

2.1.3 A Poly Peptide chain

is a chain of amino acid residues linked together by peptide bonds. The **backbone** of the polypeptide is given by the repeated sequence of three atoms of each residue in the chain: the amide N, the alpha Carbon C_{alpha} and the Carbonyl C. Rotations in the chain take place about the bonds in the backbone, whereat the peptide bond usually is inflexible. The existence of an amino group (**N-Terminal**) at one end of the chain and a carboxy group (**C-Terminal**) at the other end designs a direction to the chain. Conventionally the beginning of a polypeptides is its N-Terminal





13

A protein is a naturally occurring polypeptide with a definite 3-dimensional structure.

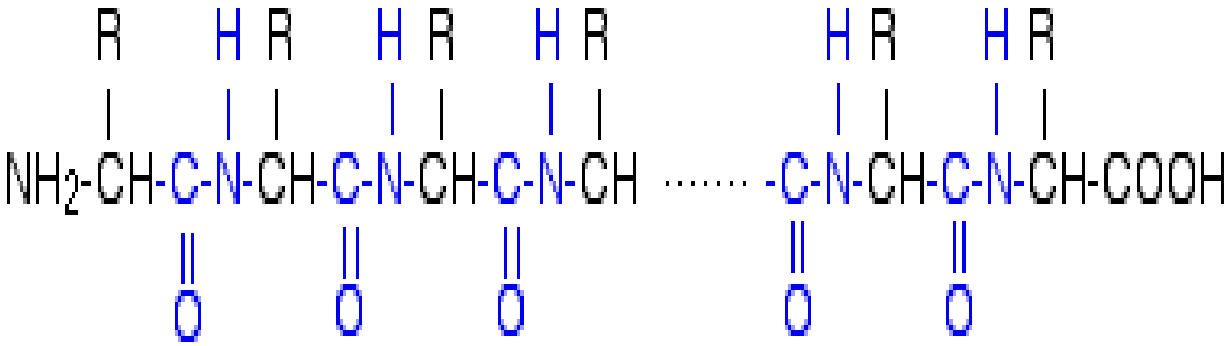
2.1.4 X – ray crystallography

Is the science of determining the arrangement of atoms within a crystal from the manner in which a beam of X-rays is scattered from the electrons within the crystal? The method produces a three-dimensional picture of the density of electrons within the crystal, from which the mean atomic positions, their chemical bonds, their disorder and sundry other information can be derived.

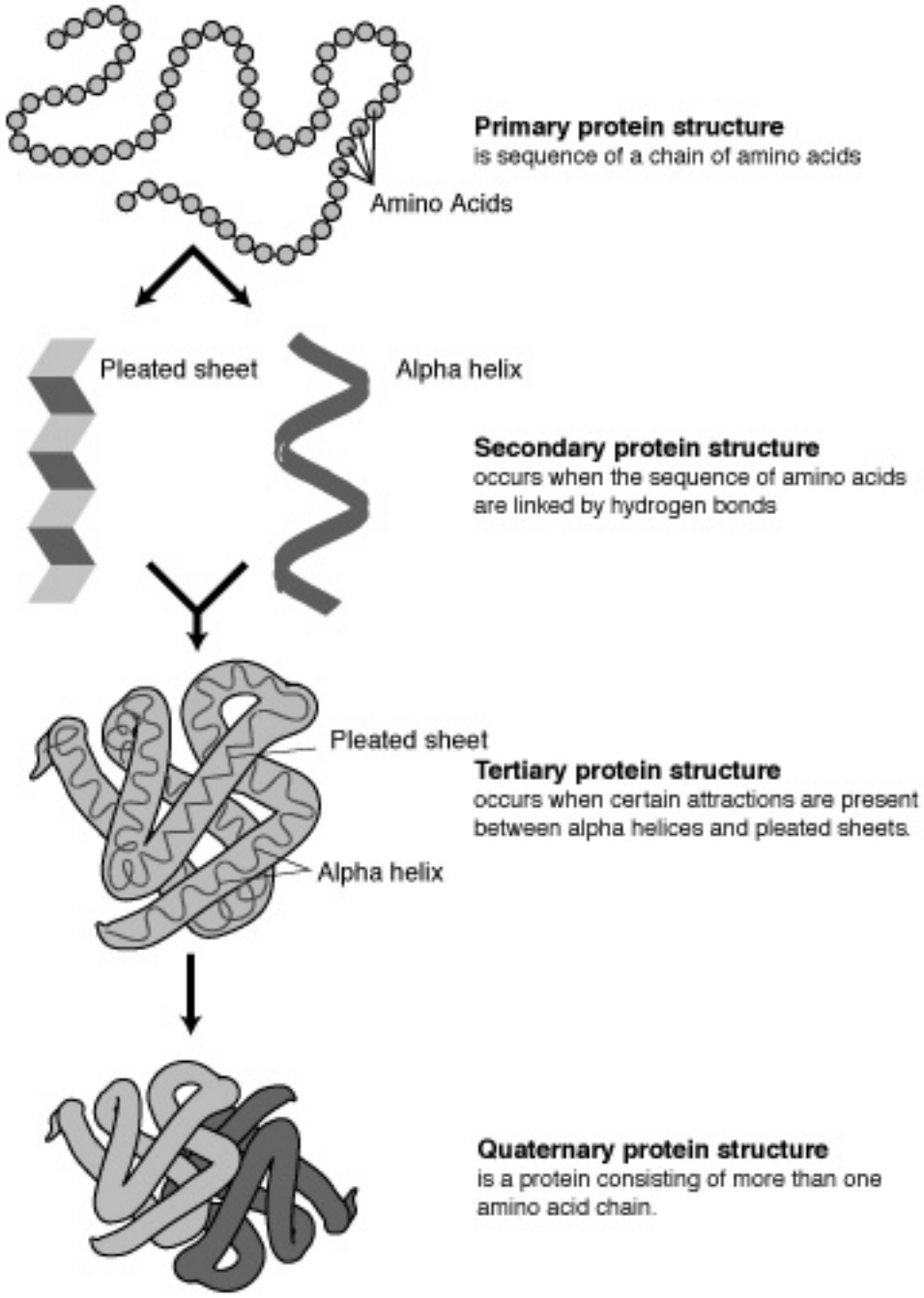
A number of residues are necessary to perform a particular biochemical function, and around 40-50 residues appears to be the lower limit for a functional domain size. Protein sizes range from this lower limit to several thousand residues in multi-functional or structural proteins. However, the current estimate for the average protein length is around 300 residues. Very large aggregates can be formed from protein subunits, for example many thousand actin molecules assemble into a collagen filament.

2.2 Protein Structure

Proteins fold in three dimensions. Protein structure is organized hierarchically from so-called *primary structure* to *quaternary structure*. Higher-level structures are *motifs* and *domains*.



14

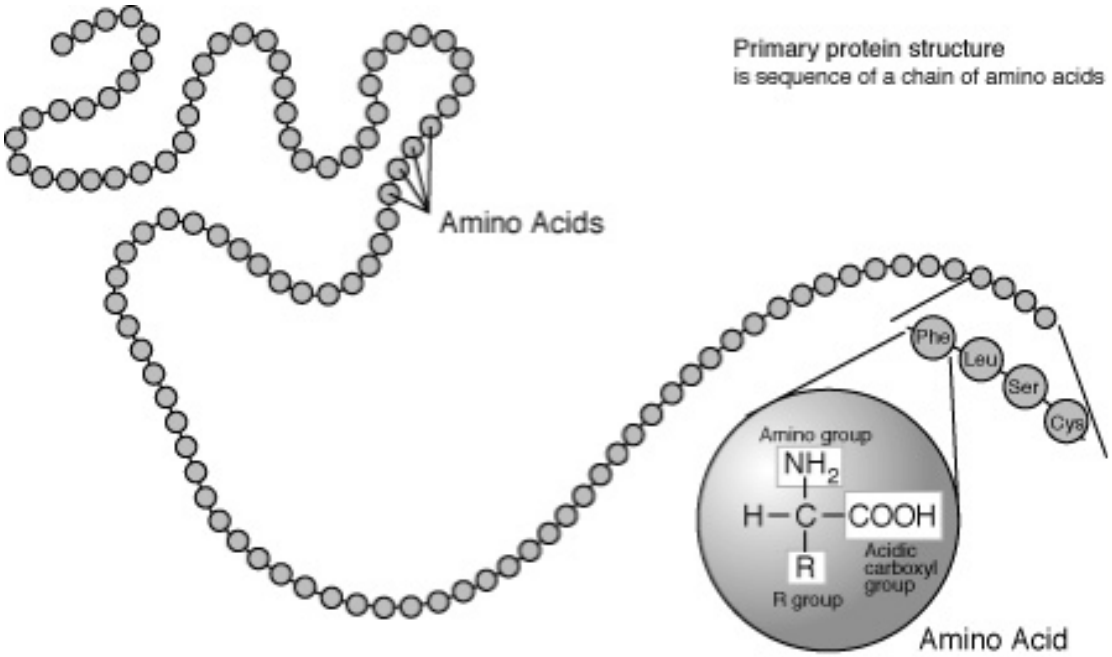


2.2.1 Primary Structure

the amino acid sequence of the peptide chains. The sequence of the different amino acids is called the primary structure of the peptide or protein. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a

15

process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code. Post-transcriptional modifications such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene.



Primary Protien Structure is a sequence of a chain of amino acids

2.2.2 Secondary Structure

Highly regular sub-structures (*alpha helix* and *strands of beta sheet*) which are locally defined, meaning that there can be many different secondary motifs present in one single protein molecule. By building models of peptides using known information about bond lengths and angles, the first elements of secondary structure, the alpha helix and the beta sheet, were suggested in 1951 by Linus Pauling and coworkers. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. These secondary structure elements only depend on properties that all the residues have in common, explaining why they occur frequently in most proteins. Since then other elements of secondary structure have been discovered such as various loops and other forms of helices. The part of the backbone that is not in a regular secondary structure is said to be

16

random coil. Each of these two secondary structure elements have a regular geometry, meaning they are constrained to specific values of the dihedral angles ψ and φ. Thus they can be found in a specific region of the Ramachandran plot. Turns, loops and a few other secondary structure

elements such as a 3-10 helix complete the picture. We have now enough pieces to assemble a complete protein, displaying its typical tertiary structure.

2.2.3 Tertiary Structure

Three-dimensional structure of a single protein molecule: a spatial arrangement of the secondary structures. It also describes the completely folded and compacted polypeptide chain. The elements of secondary structure are usually folded into a compact shape using a variety of loops and turns. The formation of tertiary structure is usually driven by the burial of hydrophobic residues, but other interactions such as hydrogen bonding, ionic interactions and disulfide bonds can also stabilize the tertiary structure. The tertiary structure encompasses all the noncovalent interactions that are not considered secondary structure, and is what defines the overall fold of the protein, and is usually indispensable for the function of the protein.

2.2.4 Quaternary Structure

Complex of several protein molecules or polypeptide chains, usually called protein subunits in this context, which function as part of the larger assembly or protein complex. The quaternary structure is the interaction between several chains of peptide bonds. The individual chains are called subunits. The individual subunits are not necessarily covalently connected, but might be connected by a disulfide bond. Not all proteins have quaternary structure, since they might be functional as monomers. The quaternary structure is stabilized by the same range of interactions as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, and a tetramer if it contains four subunits. Multimers made up of identical subunits may be referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits may be referred to with a prefix of "hetero-" (e.g. a heterodimer). Tertiary structures vary greatly from one protein to another. They are held together by glycosydic and covalent bonds.

17

2.2.5 Additional Structures

In addition to these levels of structure - a protein may shift between several similar structures in performing its biological function. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as chemical conformation, and transitions between them are called conformational changes.

The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. These peptide bonds provide rigidity to the protein. The two ends of the amino acid chain are referred to as the C-terminal end or carboxyl terminus (C-terminus) and the N-terminal end or amino terminus (N-terminus) based on the nature of the free group on each extremity.

The various types of secondary structure are defined by their patterns of hydrogen bonds between the main-chain peptide groups. However, these hydrogen bonds are generally not stable by themselves, since the water-amide hydrogen bond is generally more favorable than the amide-amide hydrogen bond. Thus, secondary structure is stable only when the local concentration of water is sufficiently low, e.g., in the molten globule or fully folded states.

Similarly, the formation of molten globules and tertiary structure is driven mainly by structurally *non-specific* interactions, such as the rough propensities of the amino acids and hydrophobic interactions. However, the tertiary structure is *fixed* only when the parts of a protein domain are locked into place by structurally *specific* interactions, such as ionic interactions (salt bridges), hydrogen bonds and the tight packing of side chains. The tertiary structure of extracellular proteins can also be stabilized by disulfide bonds, which reduce the entropy of the unfolded state; disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment.

18

2.4 Side Chain Conformation

The atoms along the side chain are named with Greek letters in Greek alphabetical order: α, β, γ, δ, ϵ and so on. C_α refers to the carbon atom closest to the carbony group of that amino acid, C_β the second closest and so on. The C_α is usually considered a part of the backbone. The dihedral angles around the bonds between these atoms are named χ1, χ2, χ3 etc. E.g. the first and second carbon atom in the side chain of lysine is named α and β, and the dihedral angle around the α-β bond is named χ1. Side chains can be in different conformations called gauche(-), trans and gauche(+). Side chains generally tend to try to come into a staggered conformation around χ2, driven by the minimization of the overlap between the electron orbitals of the hydrogen atoms.

19

Ch3: *PROTEIN DATA BANK (PDB)*

3.1 History

The Protein Data Bank (PDB) was established in 1971 as the central archive of all experimentally determined protein structure data. Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB). The mission of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community.
Founded in 1971 by Drs. Edgar Meyer and Walter Hamilton Brookhaven National Laboratory, management of the Protein Data Bank was transferred in 1998 to members of the Research Collaboratory for Structural Bioinformatics (RCSB). Rutgers University is the lead site and is currently under the direction of Helen M. Berman.

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. The founding members are RCSB PDB (USA), MSD-EBI (Europe) and PDBj (Japan). The BMRB (USA) group joined the wwPDB in 2006. The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community.

The PDB is a key resource in structural biology and is critical to more recent work in structural genomics.

Countless derived databases and projects have been developed to integrate and classify the PDB in terms of protein structure, protein function and protein evolution.
The Protein Data Bank (PDB) is a repository for 3-D structural data of proteins and nucleic acids. These data, typically obtained by X-ray crystallography or

20

NMR spectroscopy and submitted by biologists and biochemists from around the world, are released into the public domain, and can be accessed for free.

Around 90% of the protein structures available in the (Protein Data Bank) have been determined by X-ray crystallography. This method allows one to measure the 3D density distribution of electrons in the protein (in the crystallized state) and thereby infer the 3D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by Nuclear Magnetic Resonance techniques, which can also be used to determine secondary structure. Note that aspects of the secondary structure as whole can be determined via other biochemical techniques such as circular dichroism.

Secondary structure can also be predicted with a high degree of accuracy. Cryo-electron microscopy has recently become a means of determining protein structures to high resolution (less than 5 angstroms or 0.5 nanometer) and is anticipated to increase in power as a tool for high resolution work in the next decade. This technique is still a valuable resource for researchers working with very large protein complexes such as virus coat proteins and amyloid fibers.

3.1.1 Growth

When the PDB was originally founded it contained just 7 protein structures. Since then it has undergone an approximate exponential growth in the number of structures, which does not show any sign of falling off.

The growth rate of the PDB has been the subject of fairly extensive analysis.

3.1.2 Contents

As of 24 June 2008, the database contained 51,491 released atomic coordinate entries (or "structures"), 47,526 of that proteins, the rest being nucleic acids, nucleic acid-protein complexes, and a few other molecules. About 5,000 new structures are released each year. Data are stored in the mmCIF format specifically developed for the purpose. It is estimated that the size of the PDB archive will triple to 150,000 structures by the year 2014.

21

Note that the database stores information about the exact location of all atoms in a large biomolecule (although, usually without the hydrogen atoms, as their positions are more of a statistical estimate); if one is only interested in *sequence data*, i.e., the list of amino acids making up a particular protein or the list of nucleotides making up a particular nucleic acid, the much larger databases from Swiss-Prot and the International Nucleotide Sequence Database Collaboration should be used.

3.2 File Formats

Through the years the PDB file format has undergone many, many changes and revisions. Its original format was dictated by the width of computer punch cards.

PDB Format Guide - Prepared by the PDB Staff at BNL The PDB format specification can be found here, and it is vital that you read this before looking at the raw data.

-Recently PDB provides a representation of PDB data in XML format, PDBML format.

-ftp.wwpdb.org The raw data can be downloaded from here.

-PDB format files can be downloaded using HTTP with URLs like this: http://www.pdb.org/pdb/files/4hhb.pdb.gz

-PDBML (XML) files can be downloaded using HTTP with URLs like this: http://www.pdb.org/pdb/files/4hhb.xml.gz

-ftp.ebi.ac.uk/pub/databases/rcsb/ Alternate download location for the PDB archive.

-www.pdb.org Statistics about the PDB can be found here.

-This legacy format has caused many problems with the format, and consequently there are 'clean-up' projects;

-The Molecular Modeling DataBase (MMDB) from NCBI

22

-wwPDB

The MMDB uses ASN.1 (and an XML conversion of this format). The wwPDB members RCSB PDB, MSD-EBI, and PDBj are working together to make the data uniform across the archive. Some believe this to be desirable, others argue that, without a universal repository of information (i.e., a common dictionary), it is not possible to draw comparisons.

Each structure published in PDB receives a four-character alphanumeric identifier, its PDB ID. This should not be used as an identifier for biomolecules, since often several structures for the same molecule (in different environments or conformations) are contained in PDB with different PDB IDs.

If a biologist submits structure data for a protein or nucleic acid, wwPDB staff reviews and annotates the entry. The data are then automatically checked for plausibility. The source code for this validation software has been released for free. The main data base accepts only experimentally derived structures, and not theoretically predicted ones.

Various funding agencies and scientific journals now require scientists to submit their structure data to PDB.

3.3 Protein Structure Data Bases

Because the PDB releases data into the public domain, the data has been used in various other protein structure databases.

3.3.1 Examples of protein structure data bases

Database of Macromolecular Movements

describes the motions that occur in proteins and other macromolecules, particularly using movies

JenaLib

23

the Jena Library of Biological Macromolecules is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization and analysis.

MODBASE

a database of three-dimensional protein models calculated by comparative modeling

MSD

the Macromolecular Structure Database (MSD) the European project for the collection, management and distribution of data about macromolecular structures, derived in part from the PDB

OCA

a browser-database for protein structure/function - The OCA integrates information from KEGG, OMIM, PDBselect, Pfam, PubMed, SCOP, SwissProt and others.

OPM

provides spatial positions of protein three-dimensional structures with respect to the lipid bilayer.

PDB Lite

derived from OCA, PDB Lite was provided to make it as easy as possible to find and view a macromolecule within the PDB

PDBsum

provides an overview macromolecular structures in the PDB, giving schematic diagrams of the molecules in each structure and of the interactions between them

PDBTM

the Protein Data Bank of Transmembrane Proteins a selection of the PDB.

24

PDBWiki

a community annotated knowledge base of biological molecular structures

Proteopedia

the collaborative, 3D encyclopedia of proteins and other molecules. A wiki that contains a page for every entry in the PDB (>50,000 pages), with a Jmol view that highlights functional sites and ligands. Offers an easy-to-use scene-authoring tool so you don't have to learn Jmol script language to create customized molecular scenes. Custom scenes are easily attached to "green links" in descriptive text that display those scenes in Jmol.

SCOP

the Structural Classification of Proteins a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

TOPSAN

the Open Protein Structure Annotation Network a wiki designed to collect, share and distribute information about protein three-dimensional structures.

Comparison

Protein structure database comparison table.

25

Ch 4: DATA MINING

4.1 Definition

Data mining : is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data and "the science of extracting useful information from large data sets or databases. Data mining in relation to enterprise resource planning is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making.

4.2 Background

Traditionally, business analysts have performed the task of extracting useful information from recorded data, but the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been a shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools.

26

The modern technologies of computers, networks, and sensors have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become useful. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, to data.

Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities. However, abdicating control of this process from the statistician to the machine may result in false-positives or no useful results at all.

Although data mining is a relatively new term, the technology is not. For many years, businesses have used powerful computers to sift through volumes of data such as supermarket scanner data to produce market research reports (although reporting is not considered to be data mining). Continuous innovations in

computer processing power, disk storage, and statistical software are dramatically increasing the accuracy and usefulness of data analysis.

The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g., rule-based systems) and opaque in others such as neural networks. Moreover, some data-mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery.

Metadata, or data about a given data set, are often expressed in a condensed *data-minable* format, or one that facilitates the practice of data mining. Common examples include executive summaries and scientific abstracts.

Data mining relies on the use of real world data. This data is extremely vulnerable to collinearity precisely because data from the real world may have unknown interrelations. An unavoidable weakness of data

27

mining is that the critical data that may expose any relationship might have never been observed. Alternative approaches using an experiment-based approach such as Choice Modelling for human-generated data may be used. Inherent correlations are either controlled for or removed altogether through the construction of an experimental design.

Recently, there were some efforts to define a standard for data mining, for example the CRISP-DM standard for analysis processes or the Java Data-Mining Standard. Independent of these standardization efforts, freely available open-source software systems like RapidMiner and Weka have become an informal standard for defining data-mining processes.

4.3 Privacy Concerns

There are also privacy and human rights concerns associated with data mining, specifically regarding the source of the data analyzed. Data mining provides information that may be difficult to obtain otherwise. When the data collected involves individual people, there are many questions concerning privacy, legality, and ethics. In particular, data mining government or commercial data sets for national security or law enforcement purposes has raised privacy concerns.

4.4 Notable uses of Data Mining

4.4.1 Combatting Terrorism

Data mining has been cited as the method by which the U.S. Army unit Able Danger had identified the September 11, 2001 attacks leader, Mohamed Atta, and three other 9/11 hijackers as possible members of an Al Qaeda cell operating in the U.S. more than a year before the attack

It has been suggested that both the Central Intelligence Agency and the Canadian Security Intelligence Service have employed this method.

Previous data mining to stop terrorist programs under the US government include the Terrorism Information Awareness (TIA)

28

program, Computer-Assisted Passenger Prescreening System (CAPPS II), Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE), Multistate Anti-Terrorism Information Exchange (MATRIX), and the Secure Flight program Security-MSNBC. These programs have been discontinued due to controversy over whether they violate the US Constitution's 4th amendment.

4.4.2 Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called tablebases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened up. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully have the required high level of abstraction in order to be applied successfully. Instead, extensive experimentation with the tablebases, combined with an intensive study of tablebase-answers to well designed problems and with knowledge of prior art, i.e. pre-tablebase knowledge, is used to yield insightful patterns. Berlekamp in dots-and-boxes etc. and John Nunn in chess endgames are notable examples of researchers doing this work, though they were not and are not involved in tablebase generation.

4.4.3 Business

Data mining in customer relationship management applications can contribute significantly to the bottom line. Rather than contacting a prospect or customer through a call center or sending mail, only prospects that are predicted to have a high likelihood of responding to an offer are contacted. More sophisticated methods may be used to optimize across campaigns so that we can predict which channel and which offer an individual is most likely to respond to - across all potential offers. Finally, in cases where many people will take an action without an offer, uplift modeling can be used to determine which people will have the greatest increase in responding if given an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

29

Businesses employing data mining quickly see a return on investment, but also they recognize that the number of predictive models can quickly become very large. Rather than one model to predict which customers will churn, a business could build a separate model for each region and customer type. Then instead of sending an offer to all people that are likely to churn, it may only want to send offers to customers that will likely take to offer. And finally, it may also want to determine which customers are going to be profitable over a window of time and only send the offers to those that are likely to be profitable. In order to maintain this quantity of models, they need to manage model versions and move to *automated data mining*.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

Another example of data mining, often called the market basket analysis, relates to its use in retail sales. If a clothing store records the purchases of customers, a data-mining system could identify those customers who favour silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical or inexact rules may also be present within a database. In a manufacturing application, an inexact rule may state that 73% of products which have a specific defect or problem will develop a secondary problem within the next six months.

Related to an integrated-circuit production line, an example of data mining is described in the paper "Mining IC Test Data to Optimize VLSI Testing". In this paper the application of data mining and decision analysis to the problem of die-level functional test is described. Experiments mentioned in this paper demonstrate the ability of applying a system of mining historical die-test data to create a probabilistic model of patterns of die failure which are then utilized to decide in real

30

time which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products.

4.4.4 Science and engineering

In recent years, data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education, and electrical power engineering.

In the area of study on human genetics, the important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility. In lay terms, it is to find out how the changes in an individual's DNA sequence affect the risk of developing common diseases such as cancer. This is very important to help improve the diagnosis, prevention and treatment of the diseases. The data mining technique that is used to perform this task is known as multifactor dimensionality reduction.

In the area of electrical power engineering, data mining techniques have been widely used for condition monitoring of high voltage electrical equipment. The

purpose of condition monitoring is to obtain valuable information on the insulation's health status of the equipment. Data clustering such as self-organizing map (SOM) has been applied on the vibration monitoring and analysis of transformer on-load tap-changers(OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for the exact same tap position. SOM has been applied to detect abnormal conditions and to estimate the nature of the abnormalities.

Data mining techniques have also been applied for dissolved gas analysis (DGA) on power transformers. DGA, as a diagnostics for power transformer, has been available for centuries. Data mining techniques such as SOM has been applied to analyze data and to

31

determine trends which are not obvious to the standard DGA ratio techniques such as Duval Triangle.

A fourth area of application for data mining in science/engineering is within educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning and to understand the factors influencing university student retention.

Other examples of applying data mining technique applications are biomedical data facilitated by domain ontologies, mining clinical trial data, traffic analysis using SOM, et cetera.

Ch 5: WAVELET TRANSFORM

32

5.1 DEFINITION

The word wavelet is due to Morlet and Grossmann in the early 1980s. They used the French word ondelette, meaning "small wave". Soon it was transferred to English by translating "onde" into "wave", giving "wavelet".

A wavelet is a kind of mathematical function used to divide a given function or continuous-time signal into different frequency components and study each component with a resolution that matches its scale.

5.2 BACKGROUND

The fundamental idea behind wavelets is to analyze according to scale. Indeed, some researchers in the wavelet field feel that, by using wavelets, one is adopting a whole new mindset or perspective in processing data.

Wavelets are functions that satisfy certain mathematical requirements and are used in representing data or other functions. This idea is not new. Approximation using superposition of functions has existed since the early 1800's, when Joseph Fourier discovered that he could superpose sines and cosines to represent other functions. However, in wavelet analysis, the scale that we use to look at data plays a special role. Wavelet algorithms process data at different scales or resolutions. If we look at a signal with a large "window," we would notice gross features.
Similarly, if we look at a signal with a small "window," we would notice small features. The result in wavelet analysis is to see both the forest and the trees, so to speak.

This makes wavelets interesting and useful. For many decades, scientists have wanted more appropriate functions than the sines and cosines which comprise the bases of Fourier analysis, to approximate

33

choppy signals. By their definition, these functions are non-local (and stretch out to infinity). They therefore do a very poor job in approximating sharp spikes.

But with wavelet analysis, we can use approximating functions that are contained neatly in finite domains. Wavelets are well-suited for approximating data with sharp discontinuities.

The wavelet analysis procedure is to adopt a wavelet prototype function, called an analyzing wavelet or mother wavelet.

Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients. And if you further choose the best wavelets adapted to your data, or truncate the coefficients below a threshold, your data is sparsely represented.

This sparse coding makes wavelets an excellent tool in the field of data compression.

Other applied fields that are making use of wavelets include astronomy, acoustics, nuclear engineering, sub-band coding, signal and image processing, neurophysiology, music, magnetic resonance imaging, speech discrimination, optics, fractals, turbulence, earthquake-prediction, radar, human vision, and pure mathematics applications such as solving partial differential equations.

34

5.3 History

In the history of mathematics, wavelet analysis shows many different origins. Much of the work was performed in the 1930s, and, at the time, the separate efforts did not appear to be parts of a coherent theory.

Pre-1930

Before 1930, the main branch of mathematics leading to wavelets began with Joseph Fourier (1807) with his theories of frequency analysis, now often referred to as Fourier synthesis. He asserted that any.

Fourier's assertion played an essential role in the evolution of the ideas mathematicians had about the functions. He opened up the door to a new functional universe.

After 1807, by exploring the meaning of functions, Fourier series convergence, and orthogonal systems, mathematicians gradually were led from their previous notion of frequency analysis to the notion of scale analysis. That is, analyzing f(x) by creating mathematical structures that vary in scale. How? Construct a function, shift it by some amount, and change its scale. Apply that structure in approximating a signal. Now repeat the procedure. Take that basic structure, shift it, and scale it again. Apply it to the same signal to get a new approximation. And so on. It turns out that this sort of scale analysis is less sensitive to noise because it measures the average fluctuations of the signal at different scales.

The first mention of wavelets appeared in an appendix to the thesis of A. Haar (1909). One property of the Haar wavelet is that it has compact support, which means that it vanishes outside of a finite interval. Unfortunately, Haar wavelets are not continuously differentiable which somewhat limits their applications.

The 1930s

In the 1930s, several groups working independently researched the representation of functions using scale-varying basis functions.

35

Understanding the concepts of basis functions and scale-varying basis functions is the key to understanding wavelets; the sidebar next provides a short detour lesson for those interested.

By using a scale-varying basis function called the Haar basis function (more on this later) Paul Levy, a 1930s physicist, investigated Brownian motion, a type of random signal (2). He found the Haar basis function superior to the Fourier basis functions for studying small complicated details in the Brownian motion.

1960-1980

Between 1960 and 1980, the mathematicians Guido Weiss and Ronald R. Coifman studied the simplest elements of a function space, called atoms, with the goal of finding the atoms for a common function and finding the "assembly rules" that allow the reconstruction of all the elements of the function space using these atoms. In 1980, Grossman and Morlet, a physicist and an engineer, broadly defined wavelets in the context of quantum physics. These two researchers provided a way of thinking for wavelets based on physical intuition.

Post-1980

In 1985, Stephane Mallat gave wavelets an additional jump-start through his work in digital signal processing. He discovered some relationships between quadrature mirror filters, pyramid algorithms, and orthonormal wavelet bases (more on these later). Inspired in part by these results, Y. Meyer constructed the first non-trivial wavelets. Unlike the Haar wavelets, the Meyer wavelets are continuously differentiable; however they do not have compact support. A couple of years later, Ingrid Daubechies used Mallat's work to construct a set of wavelet orthonormal basis functions that are perhaps the most elegant, and have become the cornerstone of wavelet applications today.

5.4 A wavelet transform

Is the representation of a function by wavelets. The wavelets are scaled and translated copies (known as "daughter wavelets") of a finite-length or fast-decaying oscillating waveform (known as the "mother wavelet"). Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks,

36

and for accurately deconstructing and reconstructing finite, non-periodic and/or non-stationary signals.

In formal terms, this representation is a wavelet series representation of a square-integrable function with respect to either a complete, orthonormal set of basis functions, or an overcomplete set of Frame of a vector space (also known as a Riesz basis), for the Hilbert space of square integrable functions.

Wavelet transforms are classified into discrete wavelet transforms (DWTs) and continuous wavelet transforms (CWTs). Note that both DWT and CWT are of continuous-time (analog) transforms. They can be used to represent continuous-time (analog) signals. CWTs operate over every possible scale and translation whereas DWTs use a specific subset of scale and translation values or representation grid.

5.4.1 The Fast Wavelet Transform

The DWT matrix is not sparse in general, so we face the same complexity issues that we had previously faced for the discrete Fourier transform. We solve it as we did for the FFT, by factoring the DWT into a product of a few sparse matrices using self-similarity properties. The result is an algorithm that requires only order n operations to transform an n-sample vector. This is the "fast" DWT of Mallat and Daubechies.

Wavelet packets are particular linear combinations of wavelets. They form bases which retain many of the orthogonality, smoothness, and localization properties of their parent wavelets. The coefficients in the linear combinations are computed by a recursive algorithm making each newly computed wavelet packet coefficient sequence the root of its own analysis tree.

5.4.2 Adapted Waveforms

Because we have a choice among an infinite set of basis functions, we may wish to find the best basis function for a given representation of a signal. A basis of adapted waveform is the best basis function for a given signal representation. The chosen basis carries substantial information about the signal, and if the basis description is efficient (that

37

is, very few terms in the expansion are needed to represent the signal), then that signal information has been compressed.

According to Wickerhauser, some desirable properties for adapted wavelet bases are

speedy computation of inner products with the other basis functions,

speedy superposition of the basis functions,good spatial localization, so researchers can identify the position of a signal that is contributing a large component,

good frequency localization, so researchers can identify signal oscillations; and independence, so that not too many basis elements match the same portion of the signal.

For adapted waveform analysis, researchers seek a basis in which the coefficients, when rearranged in decreasing order, decrease as rapidly as possible. to measure rates of decrease, they use tools from classical harmonic analysis including calculation of information cost functions.

This is defined as the expense of storing the chosen representation. Examples of such functions include the number above a threshold, concentration, entropy, logarithm of energy, Gauss-Markov calculations, and the theoretical dimension of a sequence.

5.5 WAVELET APPLICATIONS

5.5.1 Computer and Human Vision

In the early 1980s, David Marr began work at MIT's Artificial Intelligence Laboratory on artificial vision for robots.

He is an expert on the human visual system and his goal was to learn why the first attempts to construct a robot capable of understanding its surroundings were unsuccessful.

38

Marr believed that it was important to establish scientific foundations for vision, and that while doing so, one must limit the scope of investigation by excluding everything that depends on training, culture, and so on, and focus on the mechanical or involuntary aspects of vision. This low-level vision is the part that enables us to recreate the three-dimensional organization of the physical world around us from the excitations that stimulate the retina. Marr asked the questions:

How is it possible to define the contours of objects from the variations of their light intensity?

How is it possible to sense depth?

How is movement sensed?

He then developed working algorithmic solutions to answer each of these questions.

Marr's theory was that image processing in the human visual system has a complicated hierarchical structure that involves several layers of processing. At each processing level, the retinal system provides a visual representation that scales progressively in a geometrical manner. His arguments hinged on the detection of intensity changes. He theorized that intensity changes occur at different scales in an image, so that their optimal detection requires the use of operators of different sizes. He also theorized that sudden intensity changes produce a peak or trough in the first derivative of the image. These two hypotheses require that a vision filter have two characteristics: it should be a differential operator, and it should be capable of being tuned to act at any desired scale. Marr's operator was a wavelet that today is referred to as a "Marr wavelet".

5.5.2 FBI Fingerprint Compression

Between 1924 and today, the US Federal Bureau of Investigation has collected about 30 million sets of fingerprints.

39

The archive consists mainly of inked impressions on paper cards. Facsimile scans of the impressions are distributed among law enforcement agencies, but the digitization quality is often low. Because a number of jurisdictions are experimenting with digital storage of the prints, incompatibilities between data formats have recently become a problem. This problem led to a demand in the criminal justice community for a digitization and a compression standard.

In 1993, the FBI's Criminal Justice Information Services Division developed standards for fingerprint digitization and compression in cooperation with the National Institute of Standards and Technology, Los Alamos National Laboratory, commercial vendors, and criminal justice communities.

Let's put the data storage problem in perspective. Fingerprint images are digitized at a resolution of 500 pixels per inch with 256 levels of gray-scale information per pixel. A single fingerprint is about 700,000 pixels and needs about 0.6 Mbytes to store. A pair of hands, then, requires about 6 Mbytes of storage. So digitizing the FBI's current archive would result in about 200 terabytes of data. (Notice that at today's prices of about $900 per Gbyte for hard-disk storage, the cost of storing these uncompressed images would be about a 200 million dollars.) Obviously, data compression is important to bring these numbers down.

Fig. An FBI-digitized left thumb fingerprint.

40

5.5.3 Decrease Noisy Data

In diverse fields from planetary science to molecular spectroscopy, scientists are faced with the problem of recovering a true signal from incomplete, indirect or noisy data. Can wavelets help solve this problem? The answer is certainly "yes," through a technique called wavelet shrinkage and thresholding methods, that David Donoho has worked on for several years.

The technique works in the following way. When you decompose a data set using wavelets, you use filters that act as averaging filters and others that produce details. Some of the resulting wavelet coefficients correspond to details in the data set. If the details are small, they might be omitted without substantially affecting the main features of the data set.

The idea of thresholding, then, is to set to zero all coefficients that are less than a particular threshold. These coefficients are used in an inverse wavelet transformation to reconstruct the data set. Figure 6 is a pair of "before" and "after" illustrations of a nuclear magnetic resonance (NMR) signal. The signal is transformed, thresholded and inverse-transformed.

The technique is a significant step forward in handling noisy data because the denoising is carried out without smoothing out the sharp structures. The result is cleaned-up signal that still shows important details.

The figure shown below displays an image created by Donoho of Ingrid Daubechies (an active researcher in wavelet analysis and the inventor of smooth orthonormal wavelets of compact support), and then several close-up images of her eye: an original, an image with noise added, and finally denoised image. To denoise the image, Donoho:

41

transformed the image to the wavelet domain using Coiflets with three vanishing moments,

applied a threshold at two standard deviations, and

inverse-transformed the image to the signal domain.

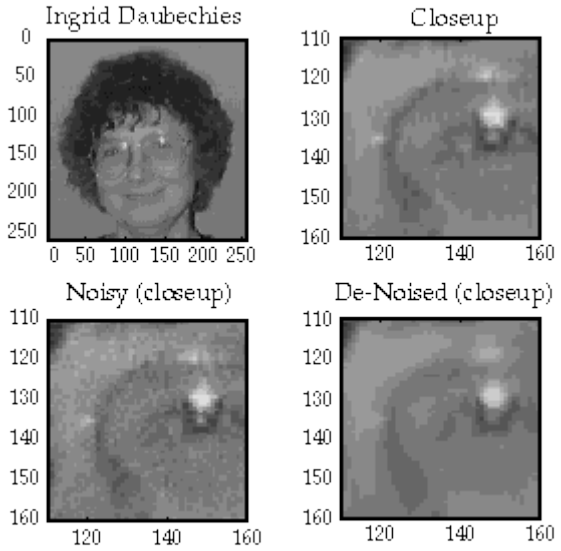


Fig. Denoising an image of Ingrid Daubechies' left eye.

5.5.4 Musical Tones

Victor Wickerhauser has suggested that wavelet packets could be useful in sound synthesis. His idea is that a single wavelet packet generator could replace a large number of oscillators. Through experimentation, a musician could determine combinations of wave packets that produce especially interesting sounds.

Wickerhauser feels that sound synthesis is a natural use of wavelets. Say one wishes to approximate the sound of a musical instrument. A sample of the notes produced by the instrument could be decomposed into its wavelet packet coefficients. Reproducing the note would then require reloading those coefficients into a wavelet packet generator and playing back the result. Transient characteristics such as attack and decay- roughly, the intensity variations of how the sound starts and ends- could be controlled separately (for example, with envelope generators), or by using longer wave packets and encoding those properties as well into each note. Any of these processes could be controlled in real time, for example, by a keyboard.

42

Notice that the musical instrument could just as well be a human voice, and the notes words or phonemes.

A wavelet-packet-based music synthesizer could store many complex sounds efficiently because

wavelet packet coefficients, like wavelet coefficients, are mostly very small for digital samples of smooth signals; and

discarding coefficients below a predetermined cutoff introduces only small errors when we are compressing the data for smooth signals.

Similarly, a wave packet-based speech synthesizer could be used to reconstruct highly compressed speech signals. Figure 8 illustrates a wavelet musical tone or toneburst.

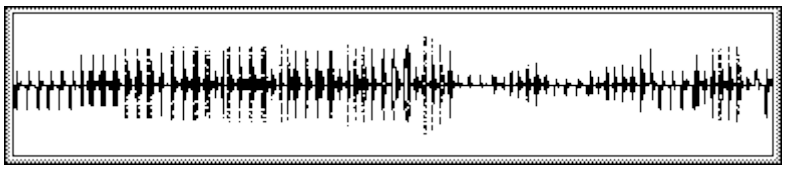


Fig.Wavelets for music: a graphical representation of a Wickerhauser toneburst.

5.6 What do Some Wavelets Look Like?

Wavelet transforms comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are.

Some of the wavelet bases have fractal structure. The Daubechies wavelet family is one example

Within each family of wavelets (such as the Daubechies family) are wavelet subclasses distinguished by the number of coefficients and by the level of iteration. Wavelets are classified within a family most often by the number of vanishing moments.

43

This is an extra set of mathematical relationships for the coefficients that must be satisfied, and is directly related to the number of coefficients (1). For example, within the Coiflet wavelet family are Coiflets with two vanishing moments, and Coiflets with three vanishing moments. In Figure 4, I illustrate several different wavelet families.

Ch 6: Filters

Wavelets are basis functions in function space. All the functions of the basis are derived from a single function , called the basic wavelet or mother wavelet, by translations and dilations. The basic wavelet in turn can be derived from a scaling function of a multiresolution analysis through the wavelet equation.

6.1 History

The oldest forms of electronic filters are passive analog linear filters, constructed using only resistors and capacitors or resistors and inductors. These are known as RC and RL single pole filters respectively. More complex multipole LC filters have also existed for many years and the operation of such filters is well understood with many books having been written about them.

Hybrid filters have also been made, typically involving combinations of analog amplifiers with mechanical resonators or delay lines. Other devices such as CCD delay lines have also been used as discrete-time filters. With the availability of digital signal processing, active digital filters have become common.

6.2 Classification by Technology

Passive filters

Passive implementations of linear filters are based on combinations of resistors (R), inductors (L) and capacitors (C). These types are

44

collectively known as passive filters, because they do not depend upon an external power supply.

Inductors block high-frequency signals and conduct low-frequency signals, while capacitors do the reverse. A filter in which the signal passes through an inductor, or in which a capacitor provides a path to earth, presents less attenuation to low-frequency signals than high-frequency signals and is a low-pass filter.

If the signal passes through a capacitor, or has a path to ground through an inductor, then the filter presents less attenuation to high-frequency signals than low-frequency signals and is a high-pass filter. Resistors on their own have no frequency-selective properties, but are added to inductors and capacitors to determine the time-constants of the circuit, and therefore the frequencies to which it responds.

At very high frequencies (above about 100 Megahertz), sometimes the inductors consist of single loops or strips of sheet metal, and the capacitors consist of adjacent strips of metal. These inductive or capacitive pieces of metal are called stubs.

The inductors and capacitors are the reactive elements of the filter. The number of elements determines the order of the filter. In this context, an LC tuned circuit being used in a band-pass or band-stop filter is considered a single element even though it consists of two components.

Active filters

Active filters are implemented using a combination of passive and active (amplifying) components, and require an outside power source. Operational amplifiers are frequently used in active filter designs. These can have high Q, and can achieve resonance without the use of inductors. However, their upper frequency limit is limited by the bandwidth of the amplifiers used.

45

Digital filters

A finite impulse response filter

Digital signal processing allows the inexpensive construction of a wide variety of filters. The signal is sampled and an analog to digital converter turns the signal into a stream of numbers. A computer program running on a CPU or a specialized DSP (or less often running on a hardware implementation of the algorithm) calculates an output number stream. This output can be converted to a signal by passing it through a digital to analog converter.

There are problems with noise introduced by the conversions, but these can be controlled and limited for many useful filters. Due to the sampling involved, the input signal must be of limited frequency content or aliasing will occur.

Other filter technologies

Quartz filters and piezoelectrics

In the late 1930s, engineers realized that small mechanical systems made of rigid materials such as quartz would acoustically resonate at radio frequencies, i.e. from audible frequencies (sound) up to several hundred megahertz.

Some early resonators were made of steel, but quartz quickly became favored. The biggest advantage of quartz is that it is piezoelectric.

This means that quartz resonators can directly convert their own mechanical motion into electrical signals. Quartz also has a very low coefficient of thermal expansion which means that quartz resonators can produce stable frequencies over a wide temperature range.

Quartz crystal filters have much higher quality factors than LCR filters. When higher stabilities are required, the crystals and their driving

46

circuits may be mounted in a "crystal oven" to control the temperature. For very narrow band filters, sometimes several crystals are operated in series.

Engineers realized that a large number of crystals could be collapsed into a single component, by mounting comb-shaped evaporations of metal on a quartz crystal. In this scheme, a "tapped delay line" reinforces the desired frequencies as the sound waves flow across the surface of the quartz crystal. The tapped delay line has become a general scheme of making high-Q filters in many different ways.

SAW filters

SAW (surface acoustic wave) filters are electromechanical devices commonly used in radio frequency applications.

Electrical signals are converted to a mechanical wave in a piezoelectric crystal; this wave is delayed as it propagates across the crystal, before being converted back to an electrical signal by further electrodes. The delayed outputs are recombined to produce a direct analog implementation of a finite impulse response filter. This hybrid filtering technique is also found in an analog sampled filter. SAW filters are limited to frequencies up to 3 GHz.

BAW filters

BAW (Bulk Acoustic Wave) filters are electromechanical devices. These filters are in the research state for the moment. BAW filters can implement ladder or lattice filters. BAW filters seem to be smaller than SAW filters, and can operate at frequencies up to 16 GHz.

6.3 Classification by Topology

Electronic filters can be classified by the technology used to implement them. Filters using passive filter and active filter technology can be further classified by the particular electronic filter topology used to implement them.

47

Any given filter transfer function may be implemented in any electronic filter topology.

Some common circuit topologies are:

Cauer topology - Passive

Sallen Key topology - Active

Multiple Feedback topology - Active

State Variable Topology - Active

Biquadratic topology biquad filter – Active

6.4 Classification by Design Methodology

Historically, linear analog filter design has evolved through three major approaches. The oldest designs are simple circuits where the main design criterion was the Q factor of the circuit. This reflected the radio receiver application of filtering as Q was a measure of the frequency selectivity of a tuning circuit. From the 1920s filters began to be designed from the image point of view, mostly being driven by the requirements of telecommunications. After WW2 the dominant methodology was network synthesis. The higher mathematics used originally required extensive tables of polynomial coefficient values to be published but modern computer resources have made that unnecessary.

6.5 Harr Filter And Adaptive Median Filter

Harr filter is used to extract the object. Moreover, nonmaxima suppression technique is adopted to get the edge localization. Finally, the adaptive

48

hysteresis thresholding is applied to get the final result. The experiments show that the proposed algorithm can be detected well-localized and thin edges.

This plugin's purpose is to remove noise from noisy images. It makes use of image reconstruction from thresholded Haar wavelet transform coefficients. Thresholding factors should be input in the dialog, and will influence the amount of noise removed. An image with the removed noise can be generated.

Finding adequate factors is critical to remove sufficient noise without affecting signal, and these factors may vary from one image type to another. Too small thresholding factors remove too little noise, too large thresholding factors induce wavelet blocks artifacts. With suitable coefficients, most of the noise is removed, while the structures are left unaffected, with their details preserved.

The wavelet filter is good at removing gaussian-type noise, while it can leave some kind of photon noise (very hot pixels for example).

Thus an option is provided in the form of an optionnal adaptive median filter. This filter will detect pixels that differ from their context by more than a given multiple of the neighborhood's standard deviation. If marked as outlying, the pixel value is replaced by the median value of the neighborhood. A suggested default value is 1.6 * sd. The idea behind this filter is that if an adequate sampling was chosen upon acquisition, no such outlying (extreme value) pixels should be found.

6.6 Endnote

Most of basic wavelet theory has been done. The mathematics have been worked out in excruciating detail and wavelet theory is now in the refinement stage. The refinement stage involves generalizations and extensions of wavelets, such as extending wavelet packet techniques.

49

The future of wavelets lies in the as-yet uncharted territory of applications. Wavelet techniques have not been thoroughly worked out in applications such as practical data analysis, where for example discretely sampled time-series data might need to be analyzed. Such applications offer exciting avenues for exploration.

Ch 7 :Approach and steps Used

The First step in generating both types of feature vectors involves converting the protein structure into a distance matrix. This process occurs in the following manner:

1- We obtain the 3D coordinates of the protein from the PDB and calculate the distance between the Cα atoms of each residue.

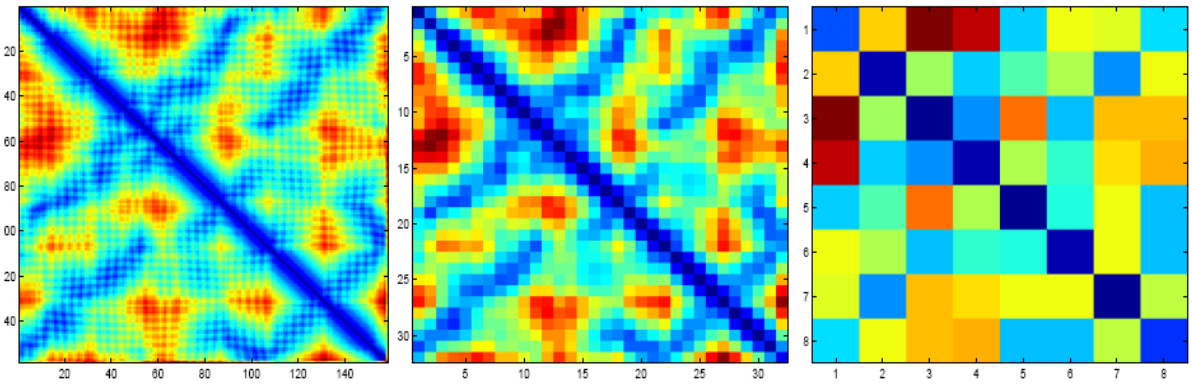
2- We place these values into an n x n matrix D, where n represents the number of residues in the protein and D (I, j) represents the distance between the Cα atoms of residues I and j. Figure 1(a) provides a graphical depiction of this matrix (for protein 1HLB), with higher elevations, or larger distances, having a lighter color.

50

3- In these matrices, secondary structures such as α-helices and parallel β-sheets emerge as dark bands along (parallel to) the main diagonal, while anti-parallel β -sheets appear perpendicular to it.
4- The image in Figure 1(a) represents a pixilated version of the distance matrix. In contrast to this discredited approach, or the related binary `contact map' representation, where distances below a certain threshold are set to 1 (0 otherwise), we operate on the actual distance values.
5- To create our global structure representation, we apply a 2D decomposition to the distance matrix. In order to use the results of this transformation as a feature vector, the final number of coefficients must be the same for every protein.
6- This can be achieved either by normalizing the size of the input (the distance matrix), or the output (the approximation coefficients). It is not immediately clear how to normalize a variable-size coefficient matrix while still preserving the necessary spatial correlations. Thus, we elect to normalize the input signal, fixing the size of the distance matrix at 128x128.This normalization occurs either through interpolation or extrapolation, depending on whether the input protein is shorter or longer than 128 residues, respectively.
7- We choose a value of 128 because discrete wavelet transformations are most effective on signals whose length is a power of 2. We prefer to interpolate and smooth, or average the excess points, over extrapolation, where we would be forced to generate additional data. Most of the proteins in our datasets are shorter than 256 residues, thus our choice of 128.
8- We perform a multi-level 2D decomposition on this normalized matrix and use the final level of approximation coefficients as our feature vector.
9- There are a fairly large number of wavelet families that can serve as filters. We tested several, but found that the simplest, the Harr, worked well for our purposes. Examples of the wavelet decomposition for protein 1HLB can be seen in Figure 1. The figure on the left illustrates the original distance matrix, while the figures in the middle and on the right correspond to the approximation coefficients produced from a 2nd and 4th level decomposition, respectively.
10- We only focus on the approximation values produced by the final level of the wavelet decomposition, so as the decomposition level increases, the number of coefficients decreases by a factor of 4. Despite this large reduction in data, important features such as secondary structures are still present in the figures. Since the matrix is symmetric

51

across the diagonal, we only need to keep the coefficients in the upper (or lower) triangle, plus those that fall on the diagonal itself.



Ch 8 : CONCLUSIONS AND FUTURE WORK

We present two methods of protein representation that allow for quick and efficient structure queries. We provide implementations for retrieving proteins based on the similarity of either their global shape or smaller substructures, an option that is not provided in any of the leading strategies. When viewed against current techniques, our method is superior in accuracy and more importantly, can answer user queries in a fraction of the time. We have tested our approach by running a number of queries on several different datasets of protein structures. We have validated our results against the labels and categorizations of a leading structural

52

database. We find that even as we progress down the hierarchy of the database, there is not a significant decrease in the accuracy of our method, indicating that the proteins we retrieve are truly correct.

In the future, we plan to test and refne our substructure matching technique. While not immediately obvious, we would like to determine whether one can obtain meaningful results if we allow substructure matching with gaps in the query sequence. In addition, we plan to evaluate the use of space flling curves as a sampling method to see how they perform compared to our combined parallel/antiparallel approach.

Thus far, we have focused only on founding similarity among solved protein structures. We do not feel that our technique is limited solely to this task, however. We believe that the methods presented here can be used on a number of different applications, particularly simulations that operate on structure-based data. Two examples include defect tracking in molecular dynamics (MD) simulations and the modeling of protein folding pathways. MD simulations are used to model the behavior of spurious atoms as they move through a lattice of a base material, often silicon. Given a set of simulation frames, we could apply our algorithm to create a sequence describing that set, which could then be compared against other sets. Being able to characterize the behavior of these defects would be a boon to those who work in the manufacture of semiconductors.

It is well-established that the amino acid sequence of a protein determines its _nal structure. What is unknown, however, is the true nature of the role that the primary structure plays in the folding process. In addition, the intermediate steps that a protein undergoes as it transitions from an unfolded chain to its final formed structure remain a mystery. On occasion, a protein will \misfold," resulting in an abnormal shape. These abnormalities can lead to diseases such as Alzheimer's, Cystic Fibrosis, and Creutzfeldt-

53

Jakob's (the human equivalent of Bovine Spongiform Encephalopathy or Mad Cow). As a result, there has been tremendous efort to understand the protein folding process through computer simulations like Folding@Home. As these programs grow in popularity, they will generate an enormous amount of simulation data. Fast and efficient characterization techniques such as the one proposed here will be crucial if there is to be any hope at processing the results of these simulations in a timely fashion.

Given a set of simulation frames, we can apply our technique on the results from di_erent proteins and cluster them based on similarity. Large clusters may be indicative of certain key points along the folding pathway. Clusters of consistently abnormal structures may give researchers a clue as to how and why proteins misfold. In addition, there may be instances where two groups of proteins share a portion of the pathway and then split before folding into their fnal shape. Such a characterization could provide insight into the evolution of protein structures and remains a rich direction of future study.

Ch 9 : References

[1] D. Aha and D. Kibler. Instance-based learning algorithms. Machine Learning, 6:37{66, 1991.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. J. of Mo. Biol., 215:403{410, 1990.

[3] S. F. Altschul, T. L. Madden, A. A. Scha_er, J. Zhang, Z. Anang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids

54

Research, 25:3389{3402, 1997.
[4] Z. Aung and K.-L. Tan. Rapid 3d protein structure database searching using information retrieval techniques. Bioinformatics, pages 1045{1052, 2004.
[5] J. L. Bentley. Multidimensional binary search trees used for associate searching. Comm. ACM, 18(9):509{517, 1975.
[6] A. Bhattacharya, T. Can, T. Kahveci, A. Singh, and Y. Wang. ProGreSS: Simultaneous searching of protein databases by sequence and structure. In Paci_c Symposium on Biocomputing, volume 9, pages 264{275. World Scienti_c Press, 2004.
[7] O. C_amo_glu, T. Kahveci, and A. Singh. Towards index-based similarity search for protein structure databases. In 2nd IEEE Computer Society Bioinformatics Conference (CSB), pages 148{158, 2003.
[8] F. Gao and M. Zaki. PSIST: Indexing protein structures using su_x trees. In IEEE Computational Systems Bioinformatics Conference, Palo Alto, CA, August 2005. IEEE.
[9] S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande. Folding@home and genome@home: Using distributed computing to tackle previously intractable problems in computational biology. In R. Grant, editor, Computational Genomics. Horizon Press, 2002.
[10] S. Mallat. A Wavelet Tour of Signal Processing. Academic, New York, 2nd edition, 1999.
[11] K. Marsolo and S. Parthasarathy. Alternate representation of distance matrices for characterization of protein structure. In 5th IEEE International Conference on Data Mining (ICDM05), 2005.
[12] S. Mehta, S. Barr, A. Choy, H. Yang, S. Parthasarathy, R. Machiraju, and J. Wilkins. Dynamic classi_cation of anomalous structures in

55

molecular dynamics simulation data. In Proceedings of the SIAM Conference on Data Mining, 2005.
[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classi_cation of proteins database for the investigation of sequences and structures. J. Mol. Biol, 247:536{540, 1995.
[14] L. Platzman and J. Bartholdi. Space_lling curves and the planar travelling salesman problem. J. Assoc. Comput. Mach, 46:719{737, 1989.
[15] S. Tata and J. Patel. PiQA: An algebra for querying protein data sets. In SSDBM, pages 141{150, 2003.

Operating System	Windows 95/98/Me/NT/2000/xp/vista
Memory	256MB RAM (512MB recommended)

56

Hard Drive	40MB free space
Video	640x480 with 256 colors (800x600 with 16M colors recommended)

Microsoft Visual Studio 2005.
Framework 2.

الحمد الله الذى بنعمته تتم الصالحات

57