

## Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

### Data/Information Visualisation

"A picture is worth a thousand words"

[naydin@yildiz.edu.tr](mailto:naydin@yildiz.edu.tr)

<http://www3.yildiz.edu.tr/~naydin>

1

## Introducing Visualisation

- Visualisation techniques are an important part of Bioinformatics.
- The increasing amounts of data, together with its associated complexity, mean that better human-data interfaces are needed.
- The days of simple, flat disk-files printed to an 80 times 40 character terminal are long gone.
- In the modern Bioinformatics world, much more effective presentation systems are required and they usually need to provide extra interactive capabilities.

2

## Visualize in dictionaries

- American Heritage dictionary
  - To form a mental image of
- Merriam-Webster
  - To form a mental visual image
- Concise Oxford dictionary
  - Form a mental image of
- Cambridge Dictionary
  - To form a picture of someone or something in your mind, in order to imagine or remember him, her, or it

3

## Introducing Visualisation

- **Visualisation** means the generation and presentation of pictures that help people understand a particular feature of a dataset, making data mean something to people.
  - This definition of visualisation might be a little too specific
    - as pictures (visions or graphics) are only one of a number of ways of representing information
- a graphical representation of data or concepts
- **Visualisation** is the most widely used of a general class of **perception** technologies.
  - Perception:
    - The ability to see, hear, or become aware of something through the senses
    - The way in which something is regarded, understood, or interpreted.

4

## Introducing Visualisation

- "A method of computer science to transform the symbolic into the geometric, to form a mental model and foster unexpected insights (McCormick et al., 1987)
- "... finding the artificial memory that best supports our natural means of perception." (Bertin, 1983)
- The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system.

5

## Introducing Visualisation

- There is no conceptual reason why the other human senses - **smell**, **taste**, **touch** or **hearing** - cannot be used as representations of biological data.
- However, there are some very practical limitations:
  - The need for an appropriate method of representation.
    - How would you hear, taste or feel representations of biological data?
    - Could you hear DNA intron splicing sites or smell Microarray clustering results?
      - Maybe people can, but the success of the use of visual graphics has discouraged the use of the other senses for representation of data

6

## Introducing Visualisation

- The "bandwidth" of the other senses has an effect.
  - In this context, bandwidth refers to the amount of information that can be communicated per unit time.
- The problem with using the other senses is that the vision system in humans is so highly developed that it has an extremely large processing capability in comparison to the other senses.
- Probably the closest is hearing.
  - However, even this is a poor substitute when compared to vision.

7

## Introducing Visualisation

- Consider, too, the motivation behind presenting data:
  - In many cases, it is to summarize information and provide for the identification of patterns.
- The human visual system excels at pattern recognition as there has been constant pressure, over millions of years, to select for this ability.
  - For humans and their ancestors a good vision system was essential for survival, both for finding food and for avoiding becoming something else's food.
  - Even in the modern technological world, this is still useful for practical activities such as avoiding cars while out shopping.
- This very same visual system is also excellent for analysing abstract diagrams derived from biological datasets, hence the popularity of visual representations within the biosciences .

8

## Introducing Visualisation

- Why is visualisation so important?
  - The answer is "because people are".
    - People ultimately drive science, make it what it is, as well as shape what it will become.
      - In the real world, it is people alone who are truly creative, it is people who know when they have an idea, and it is people who have the resources for scientific development and people allocate them.
  - Clever algorithms and integrated databases are no more than useful tools for people who understand how to use them (and know what they are doing).
  - Visualisation technologies help researchers gain insight into the world because they present data and information in a way that is meaningful to humans.
  - In a similar (and somewhat loose) sense, statistics have the same function in computational numerical analysis.

9

## The advantages of visualization

- Visualization provides an ability to comprehend huge amounts of data.
- Visualization allows the perception of emergent properties that were not anticipated.
  - The perception of a pattern can often be the basis of a new insight.
- Visualization often enables problems with the data to become immediately apparent.
  - A visualization commonly reveals things not only about the data itself but also about the way it is collected.

10

## The advantages of visualization

- With an appropriate visualization, errors and artifacts in the data often jump out at you.
- For this reason, visualizations can be invaluable in quality control.
- Visualization facilitates understanding of both large-scale and small-scale features of the data.
  - It can be especially valuable in allowing the perception of patterns linking local features.
- Visualization facilitates hypothesis formation.

11

## Data/Information Visualisation

- Data visualization is the presentation of data in a pictorial or graphical format.
  - It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.
- Information visualization is the study of visual representations of abstract data to reinforce human cognition;
  - hence, it is very important for decision making.

12

## Data/Information Visualisation

- A **data visualization** is a graphical representation of **quantifiable data**, usually by means of well-known chart, graph or map types.
  - Although they can be created by hand, they can always be generated by applying automated methods on top of the data.
- An **information visualization (infographic)** is a graphical representation that combines one or more data visualizations with other **non-data elements** - such as graphics or text - to point out relationships, show a process or tell a story that cannot be automatically discerned from the data alone.
  - An **infographic** requires the application of a creative process with some understanding of the underlying data and its context.

13

## Data/Information Visualisation

- The grand challenge is to focus not simply on **computational methods** of displaying large quantities of data but on both **perception** and **cognition** of such large amounts of data.
- One aspect is to focus on how the process of **computer visualization** can be improved to mirror the process of **natural visualization**.
  - Our perceptual systems were designed specifically for survival in and understanding of the surrounding external environment,
    - not abstract objects and images

14

## Data/Information Visualisation

- **Visual analysis** is becoming an essential component of medical visualization due to the rapidly growing role and availability of complex multidimensional, time-varying, mixed-modality, simulation, and multisubject datasets.
- The magnitude, complexity, and heterogeneity of the data necessitate the use of **visual analysis** techniques for diagnosis and medical research and, even more importantly, treatment planning and evaluation,
  - e.g., radiotherapy planning and post-chemotherapy evaluation

15

## Purposes of Information Visualization

- To help:
  - **Explore/Calculate**
    - Analyze
    - Reason about Information
  - **Communicate**
    - Explain
    - Make Decisions
    - Reason about Information
  - **Decorate**

16

## Goals of Information Visualization

- More specifically, visualization should:
  - **Make large datasets coherent**
    - Present huge amounts of information compactly
  - **Present information from various viewpoints**
  - **Present information at several levels of detail**
    - from overviews to fine structure
  - **Support visual comparisons**
  - **Tell stories about the data**

17

## Why Visualization?

- Use the eye for pattern recognition;
  - **people are good at**
    - scanning
    - recognizing
    - remembering images
- Graphical elements facilitate comparisons via
  - length
  - shape
  - orientation
  - texture
- Animation shows changes across time
- **Color** helps make distinctions
- Aesthetics make the process appealing

18

## A Key Question

- How do we
  - Convert abstract information into a visual representation
    - While still preserving the underlying meaning
    - And at the same time providing new insight?

19

## Fundamentals of Visualization

- Verbal Information vs Visual Information
  - Letting aside olfactory information (smell), gustatory information (taste), and haptic information (touch), and following the dual-coding theory (Paivio and Csapo 1973) we separate visual information (images) and verbal information (spoken and written natural language).
    - [Dual-coding theory is a theory of cognition according to which humans process and represent verbal and non-verbal information in separate, related systems. For example, the brain uses a different kind of representation for the word "tree" than it does for the image of a tree.]

21

## Fundamentals of Visualization

- The most difficult problem:
  - the semantic ambiguity of our natural language
- This poses a grand challenge to computational approaches, because Von-Neumann machines are missing the context!
- A computer does not know that noses can run and feet can smell
- However, one solution lies in machine learning approaches where we can train the machines to learn the context.



22

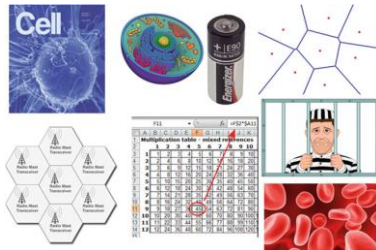
## Fundamentals of Visualization

- In the next slide we see the title page of the Journal Cell (www.cell.com):
  - the Latin letters C-e-l-l describe the basic structural, functional, and biological unit of all known living organisms.
    - Cells are the smallest unit of life and therefore we may call it the basic building block of life.
- However, a huge problem with verbal information is that we are confronted with semantic ambiguity,
  - which means that a word has often more than one meaning

23

## Fundamentals of Visualization

- CELL
  - Which one?



24

## Fundamentals of Visualization

- The word cell has a lot of different meanings:
  - the famous Journal,
  - the basic building block of life,
  - a battery cell,
  - a Voronoi cell in mathematical topology,
  - a prisoner's cell,
  - a cell of a radio network,
  - a blood cell,
  - a cell of a spreadsheet,
  - a cell in aircrafts or car manufacturing,
  - a foam cell,
  - ...
- For a better understanding, it is useful to review more detailed the already learned human information processing.

25

## Visual Information Processing (Pictures)

- According to the theory of Multimedia Learning by Mayer (2001) perceived physical visual stimuli (e.g., images) are pictorially processed and are thus cognitive “similar” to the original physical real-world data.
  - Pictures are physically perceived intentionally by the eyes and then briefly hold in the so-called visual sensory register.
  - Only if there is a certain amount of attention the pictures will become represented within the working memory.

26

## Visual Information Processing (Pictures)

- Once the working memory is full of image pieces (cognitive overload), the next active cognitive processing involves organizing those pieces into a coherent structure.
- The resulting knowledge representation is a pictorial model,
  - that is, the person builds an organized visual representation of the main parts of the picture.
- Finally, active cognitive processing is required to connect this new representation with previous knowledge.
- If this happens then the picture will be memorized in the long-time memory

27

## Verbal Information Processing: Written Text

- Written text, i.e., written natural language (words) are perceived as images, but processed symbolically as text.
- Consequently our “natural” language is a kind of artificial concept to represent real-world data:
  - The presentation of printed text creates an information-processing challenge for the dual-channel system:
    - Although the words are presented visually so they are initially perceived through the eyes and thus brought into the working memory as image, they must be mentally processed by the auditory part of the working memory, thus processed like spoken words.

28

## Verbal Information Processing: Written Text

- Consequently, when verbal material must enter through the visual channel, the words must take a complex route through the system, and must also compete for attention with images which might be perceived in parallel through the visual channel.
- The consequences of this problem are addressed in the modality principle (Moreno and Mayer 1999).
  - The modality principle states that
    - low-experience learners more successfully understand information that uses narration rather than on-screen text.

29

## Verbal Information Processing: Spoken Text

- Natural language (text) can also be perceived directly as spoken words, consequently directly auditory processed:
  - In this case the piece of text (word) is picked up as sound by the ears and held temporarily in the auditory sensory memory.
- If the person pays attention to the sounds coming into the ears, some of the incoming sounds will be selected for inclusion in the so-called word sound base.
- The words in the word base are disorganized fragments, so the next step is to build them into a coherent mental structure.
  - In this process, the words change from being represented based on sound to being represented based on word meaning.
- The person may use prior knowledge to integrate the new words into the word base—this is ensuring the context—
  - However, this is missing in our computational approaches so far, but future advances in machine learning may bring a significant step forwards, because these approaches “learn” from the environment.

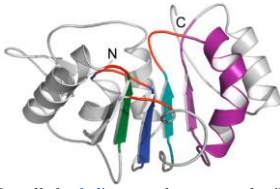
30

## Fundamentals of Visualization

- Is a Picture Really Worth a Thousand Words?
  - The answer is: it depends!
    - Some researchers are arguing that sometimes a picture might be worth a billion words (Michel et al., 2011),
    - whereas others are arguing that sometimes text is better than an image.
- This famous proverb refers to the concept that a complex idea can be conveyed with just one single image and infers a central goal of visualization:
  - to make it possible to perceive and cognitively process large amounts of data quickly.
- The following image is a good example on how a picture can explain a complex idea:
  - A ribbon diagram aka Richardson diagram, (Richardson, 2000), is a standard method of schematic protein representation.

31

## Fundamentals of Visualization



- The ribbon shows the overall path and organization of the protein backbone and is generated by interpolating a smooth curve through the polypeptide backbone.
- So-called  $\alpha$ -helices are shown as curly ribbons,  $\beta$ -strands as arrows, and thin lines for non-repetitive coils or loops.
- The direction of the polypeptide chain is shown locally by the arrows, and may be indicated overall by a color ramp along the length of the ribbon.
- Such diagrams are useful for expressing the molecular structure (twist, fold, and unfold)

32

## Informatics as Semiotics Engineering

- **Semiotics** is the study of signs and symbols as a significant part of communication.
  - As different from linguistics, however, semiotics also studies non-linguistic sign systems.
- **Semiotics** is often divided into three branches:
  - **Semantics:**
    - relation between signs and the things to which they refer; their signified meaning
  - **Syntactics:**
    - relations among or between signs in formal structures
  - **Pragmatics:**
    - relation between signs and sign-using agents or interpreters

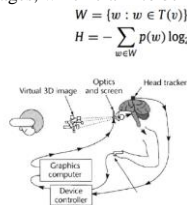
33

## Informatics as Semiotics Engineering

- Three examples of languages, which claim to be visual:



Ware, C. (2004) *Information Visualization: Perception for Design (Interactive Technologies) 2nd Edition*. San Francisco, Morgan Kaufmann.



$$W = \{w : w \in T(v)\}$$

$$H = - \sum_{w \in W} p(w) \log_2 p(w)$$

- Cave paintings (= images), which can be directly interpreted;
- Schematic diagram, showing a virtual environment and the human-computer interaction on a fairly abstract level;
- Expression of a mathematical equation, that is on a highly abstract level;

34

## Informatics as Semiotics Engineering

- Computer Science lacks a reliable concept of the human mind, whereas the psychological science lacks solid concepts for algorithms and data structures;
  - consequently, there is a need for a theory in which both domains find a place (Andersen, 2001).
- A sign integrates two sides:
  - physical (=signifier)
  - psychological (=signified).
- **Semiotics** is the study of signs and therefore can talk about representations
  - (algorithms and data structures as signifiers) and the interpretation by the end user (domain concepts as the signified).
- However, only those parts of the computational processes that influence the interpretation, and only those parts of the interpretations that are influenced by the computation, may be analyzed by semiotic methods

35

## Some Definitions

- **Visualization** is a method of computer science
  - to transform the symbolic into the geometric,
  - to support the formation of a mental model and foster insights;
    - as such it is an essential component of the knowledge discovery process.
- **Information visualization** is the interdisciplinary study of the visual representation of large-scale collections of non-numerical data,
  - such as files and software, databases, networks, etc.,
- to allow users to see, explore, and understand information at once.

36

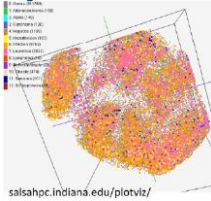
## Some Definitions

- **Data visualization** is the visual representation of complex data,
  - to communicate information clearly and effectively, making data useful and usable.
- **Visual Analytics** focuses on analytical reasoning of complex data facilitated by interactive visual interfaces.
- **Content Analytics** is a general term addressing so-called **unstructured data**—mainly text—
  - by using mixed methods from visual analytics and business intelligence.

37

## The Visualization Process

- Visualization is a typical HCI topic.
  - Large-scale high dimensional data visualization is highly valuable for scientific discovery in many fields of data mining and information retrieval.

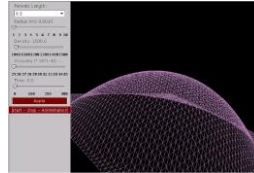


- PlotViz is a 3D data point browser that visualizes large volume of 2- or 3-dimensional data as points in a virtual space on a computer screen and enable users to explore the virtual space interactively.

38

## The Visualization Process

- Interactive visualizations provide the ability to comprehend data and to interactively analyze information properties.



- This example is about an interactive visualization to enhance student understanding of complex data.
- Simulations are assumed to offer various benefits, especially to novice

- medical students learning theoretical concepts, processes, relationships, as well as invasive procedural skills, which is extremely important within decreasing clinical exposure.
- Consequently, students can acquire knowledge in a safe environment

39

## The Visualization Process

- The process of data visualization includes four basic stages, combined in a number of feedback loops:
  - The collection and storage of data.
  - A preprocessing stage designed to transform the data into something that is easier to manipulate.
    - Usually there is some form of data reduction to reveal selected aspects.
    - Data exploration is the process of changing the subset that is currently being viewed.

40

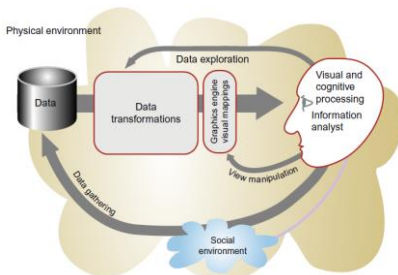
## The Visualization Process

- Mapping from the selected data to a visual representation,
  - which is accomplished through computer algorithms that produce an image on the screen.
    - User input can transform the mappings, highlight subsets, or transform the view.
      - Generally this is done on the user's own computer.
  - The human perceptual and cognitive system (the perceiver).

41

## The Visualization Process

- Taking all these considerations into account interactive visualization is a typical human-computer interaction task:



42

## The Visualization Process

- Ward et al. (2010) follow the notion that there is no distinction between data and information visualization
  - both provide representations of data;
- However, the datasets might be different.
- To interactively making the data understandable for the end user is a typical task from HCI (Holzinger 2013).

43





## A Taxonomy of Visualization Methods

- **Concept Visualization**, such as a **concept map** or a **Gantt chart**;
- these are methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses through the help of rule-guided mapping procedures.
  - In concept visualization knowledge is usually presented in a 2D graphical display where concepts (usually represented within boxes or circles), connected by directed arcs encoding brief relationships (linking phrases) between pairs of concepts.
  - These relationships usually consist of verbs, forming propositions or phrases for each pair of concepts.

50

## A Taxonomy of Visualization Methods

- **Metaphor Visualization** such as **metro maps** or **story template** can be used as effective and simple templates to convey complex insights.
- Visual Metaphors fulfill a dual function:
  - First, they position information graphically to organize and structure it.
  - Second, they convey an insight about the represented information through the key characteristics of the metaphor that is employed.

51

## A Taxonomy of Visualization Methods

- **Compound Visualization** consists of several of the aforementioned formats.
  - They can be complex knowledge maps that contain diagrammatic and metaphoric elements, conceptual cartoons with quantitative charts, or wall sized infomurals.
    - **infomural**: A visually engaging representation of your journey
  - This label thus typically designates the complementary use of different graphic representation formats in one single schema or frame.
  - They result from two (or more) spatially distinct different data representations, each of which can operate independently,
    - but can be used together to correlate information in one representation with that in another.

52

## Visual Principles

- Types of Graphs
- Pre-attentive Properties
- Relative Expressiveness of Visual Cues
- Visual Illusions
- Tufte's notions
  - Graphical Excellence
  - Data-Ink Ratio Maximization
  - How to Lie with Visualization





53

## A Graph

- A visual display that illustrates one or more relationships among entities
- A shorthand way to present information
  - Allows a **trend**, **pattern**, or **comparison** to be easily apprehended
- Anatomy of a Graph
  - **Framework**
    - sets the stage
    - kinds of measurements, scale, ...
  - **Content**
    - marks
    - point symbols, lines, areas, bars, ...
  - **Labels**
    - title, axes, tic marks, ...

54

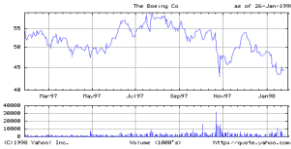
## Types of Symbolic Displays

- **Graphs** → 
- **Charts** → 
- **Maps** → 
- **Diagrams** → 

55

# Types of Symbolic Displays

- Graphs
  - at least two scales required
  - values associated by a symmetric “paired with” relation
  - Examples: scatter-plot, bar-chart, layer-graph

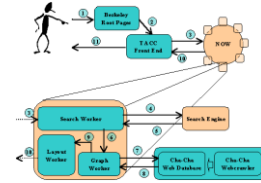


56

# Types of Symbolic Displays

- Charts
  - discrete relations among discrete entities
  - structure relates entities to one another
  - lines and relative position serve as links

Examples:  
family tree  
flow chart  
network diagram



57

# Types of Symbolic Displays

- Maps
  - internal relations determined (in part) by the spatial relations of what is pictured
  - labels paired with locations

Examples:  
map of census data  
topographic maps  
From www.thehighsierra.com

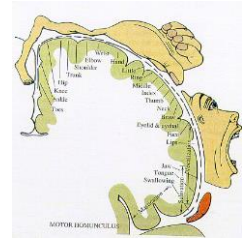


58

# Types of Symbolic Displays

- Diagrams
  - schematic pictures of objects or entities
  - parts are symbolic (unlike photographs)
    - how-to illustrations
    - figures in a manual

From Gletman, Henry. Psychology. W.W. Norton and Company, Inc. New York, 1995



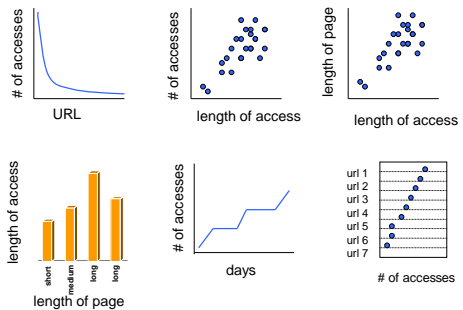
59

# Basic Types of Data

- Nominal (qualitative)
  - (no inherent order)
  - city names, types of diseases, ...
- Ordinal (qualitative)
  - (ordered, but not at measurable intervals)
  - first, second, third, ...
  - cold, warm, hot
- Interval (quantitative)
  - list of integers or reals

60

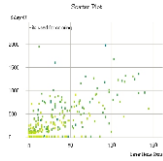
# Common Graph Types



61

## Scatter Plots

- Qualitatively determine if variables



- are highly correlated
  - linear mapping between horizontal & vertical axes
- have low correlation
  - spherical, rectangular, or irregular distributions

- have a nonlinear relationship
  - a curvature in the pattern of plotted points

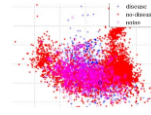
- Place points of interest in context

- color representing special entities

62

## Visualizations for Multivariate Data

- Some important visualization methods:

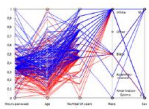


- Scatterplots are the oldest, point-based techniques, and projects (maps) data from an n-dimensional space into an arbitrary k-dimensional display space

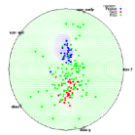
- To verify cluster separation in high dimensional data, analysts often reduce the data with a dimension reduction technique, and then visualize it with 2D Scatterplots, interactive 3D Scatterplots, or Scatterplot Matrices

63

## Visualizations for Multivariate Data



- Parallel Coordinates (PCP) is best suited for the study of high dimensional geometry, where each data point is plotted as a polyline.

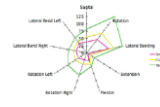


- Radial Coordinate Visualization (RadViz) is a “force-driven” point layout technique, based on Hooke’s law for equilibrium.

- A visual survey of visualization techniques for time-oriented data can be found here: <http://survey.timeviz.net>.

64

## Visualizations for Multivariate Data



- Radar Chart (star plot, spider web, polar graph, polygon plot) is a radial axis technique.



- Heatmap is tabular display technique using color instead of figures for the entities.



- Glyph is a visual representation of the entity, where its attributes are controlled by data attributes.



- Chernoff face is a face glyph which displays multivariate data in the shape of a human face.

65

## When to use which type?

- Line graph
  - x-axis requires quantitative variable
  - Variables have contiguous values
  - familiar/conventional ordering among ordinals
- Bar graph
  - comparison of relative point values
- Scatter plot
  - convey overall impression of relationship between two variables
- Pie Chart?
  - Emphasizing differences in proportion among a few numbers

66

## Experimentally Motivated Classification

- Graphs
- Tables (numerical)
- Tables (graphical)
- Charts (time)
- Charts (network)
- Diagrams (structure)
- Diagrams (network)
- Maps
- Cartograms
- Icons
- Pictures

67

## Interesting Findings

- Photorealistic images were least informative
  - Echoes results in icon studies – better to use less complex, more schematic images
- Graphs and tables are the most self-similar categories
  - Results in the literature comparing these are inconclusive
- Cartograms were hard to understand
  - Echoes other results – better to put points into a framed rectangle to aid spatial perception
- Temporal data more difficult to show than cyclic data
  - Recommend using animation for temporal data

68

## Visual Properties

- Preattentive Processing
- Accuracy of Interpretation of Visual Properties
- Illusions and the Relation to Graphical Integrity

[All Preattentive Processing figures from Healey 97](http://www.csc.ncsu.edu/faculty/healey/PP/PP.html)  
<http://www.csc.ncsu.edu/faculty/healey/PP/PP.html>

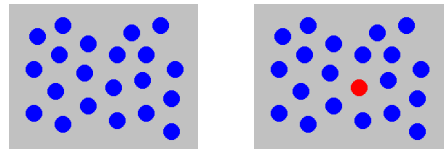
69

## Preattentive Processing

- A limited set of visual properties are processed preattentively
  - (without need for focusing attention).
- This is important for design of visualizations
  - what can be perceived immediately
  - what properties are good discriminators
  - what can mislead viewers

70

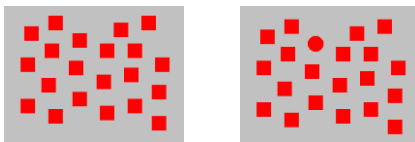
## Example: Color Selection



- Viewer can rapidly and accurately determine whether the target (red circle) is present or absent.
- Difference detected in color.

71

## Example: Shape Selection



- Viewer can rapidly and accurately determine whether the target (red circle) is present or absent.
- Difference detected in form (curvature)

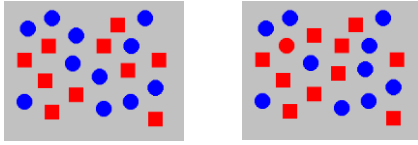
72

## Pre-attentive Processing

- < 200 - 250ms qualifies as pre-attentive
  - eye movements take at least 200ms
  - yet certain processing can be done very quickly, implying low-level processing in parallel
- If a decision takes a fixed amount of time regardless of the number of distractors,
  - it is considered to be preattentive.

73

## Example: Conjunction of Features

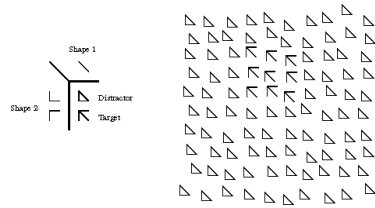


- Viewer **cannot** rapidly and accurately determine whether the target (red circle) is present or absent when target has two or more features, each of which are present in the distractors.
- Viewer **must** search sequentially.

All Preattentive Processing figures from Healey 97  
<http://www.csc.ncsu.edu/faculty/healey/PP/PP.html>

74

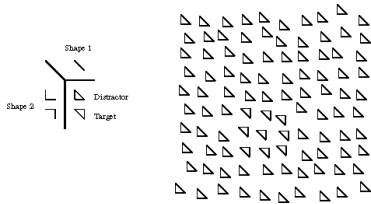
## Example: Emergent Features



- Target has a unique feature with respect to distractors (open sides)
- so the group can be detected preattentively.

75

## Example: Emergent Features

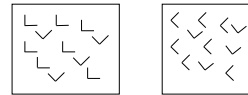


- Target does **not** have a unique feature with respect to distractors and
- so the group **cannot** be detected preattentively.

76

## Asymmetric and Graded Preattentive Properties

- Some properties are asymmetric
  - a sloped line among vertical lines is preattentive
  - a vertical line among sloped ones is not
- Some properties have a gradation
  - some more easily discriminated among than others



77

## Text NOT Preattentive

SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO  
 CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM  
 SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSLIGNE SDRO CER SNMULOC  
 GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM  
 CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM  
 GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM  
 SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSLIGNE SDRO CER SNMULOC  
 SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO  
 CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM  
 SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSLIGNE SDRO CER SNMULOC

78

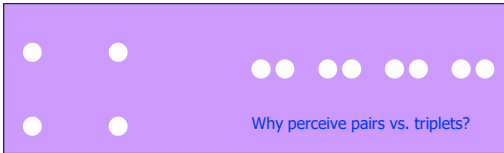
## Preattentive Visual Properties (Healey 97)

|                     |  |
|---------------------|--|
| length              | Triesman & Gormican [1988]   |
| width               | Julesz [1985]  |
| size                | Triesman & Gelade [1980]   |
| curvature           | Triesman & Gormican [1988]   |
| number              | Julesz [1985]; Trick & Pylyshyn [1994]   |
| terminators         | Julesz & Bergen [1983]   |
| intersection        | Julesz & Bergen [1983]   |
| closure             | Enns [1986]; Triesman & Souther [1985]   |
| colour (hue)        | Nagy & Sanchez [1990, 1992]; DZmura [1991]<br>Kawai et al. [1995]; Bauer et al. [1996] |
| intensity           | Beck et al. [1983]; Triesman & Gormican [1988]   |
| flicker             | Julesz [1971]  |
| direction of motion | Nakayama & Silverman [1986]; Driver & McLeod [1992]                                    |
| binocular lustre    | Wolfe & Franzel [1988]   |
| stereoscopic depth  | Nakayama & Silverman [1986]  |
| 3-D depth cues      | Enns [1990]  |
| lighting direction  | Enns [1990]  |

79

## Gestalt Properties

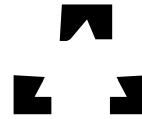
- **Gestalt:** form or configuration
  - Idea: forms or patterns transcend the stimuli used to create them.
    - Why do patterns emerge?
    - Under what circumstances?



80

## Gestalt Laws of Perceptual Organization (Kaufman 74)

- **Figure and Ground**
  - Escher illustrations are good examples
  - Vase/Face contrast
- **Subjective Contour**



81

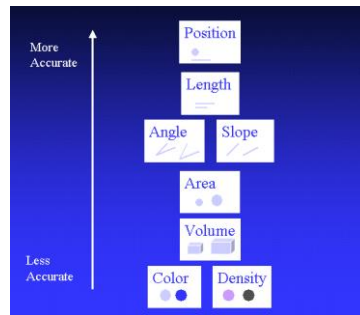
## More Gestalt Laws

- **Law of Proximity**
  - Stimulus elements that are close together will be perceived as a group
- **Law of Similarity**
  - like the preattentive processing examples
- **Law of Common Fate**
  - like preattentive motion property
    - move a subset of objects among similar ones and they will be perceived as a group

82

## Which Properties are Appropriate for Which Information Types?

- Accuracy Ranking of Quantitative Perceptual Tasks Estimated;



– only pairwise comparisons have been validated (Mackinlay 88 from Cleveland & McGill)

83

## Interpretations of Visual Properties

- Some properties can be discriminated more accurately but don't have intrinsic meaning (Senay & Ingatious 97, Kosslyn, others)
  - **Density (Greyscale)**
    - Darker -> More
  - **Size / Length / Area**
    - Larger -> More
  - **Position**
    - Leftmost -> first,
    - Topmost -> first
  - **Hue**
    - ??? no intrinsic meaning
  - **Slope**
    - ??? no intrinsic meaning

84

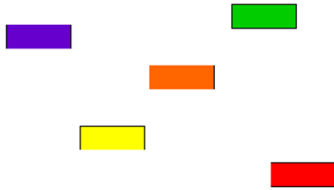
## Ranking of Applicability of Properties for Different Data Types (Mackinlay 88, Not Empirically Verified)

| QUANTITATIVE     | ORDINAL          | NOMINAL          |
|------------------|------------------|------------------|
| Position         | Position         | Position         |
| Length           | Density          | Color Hue        |
| Angle            | Color Saturation | Texture          |
| Slope            | Color Hue        | Connection       |
| Area             | Texture          | Containment      |
| Volume           | Connection       | Density          |
| Density          | Containment      | Color Saturation |
| Color Saturation | Length           | Shape            |
| Color Hue        | Angle            | Length           |

85

## Color Schemes

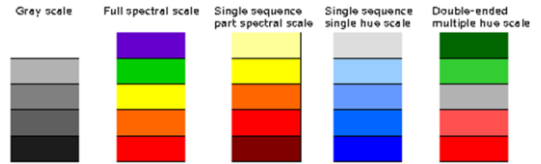
- Order these colors (low → high)



86

## Color Schemes

- Ordering examples



87

## Color Purposes

- Call attention to specific items
- Distinguish between classes of items
  - Increases the number of dimensions for encoding
- Increase the appeal of the visualization

88

## Using Color

- Proceed with caution
  - Less is more
  - Representing magnitude is tricky
- Examples
  - Red-orange-yellow-white (order)
    - Works for costs
      - Maybe because people are very experienced at reasoning shrewdly according to cost
  - Green-light green-light brown-dark brown-grey-white
    - works for atlases
  - Grayscale is unambiguous but has limited range

89

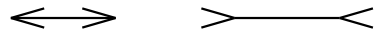
## Visual Illusions

- People don't perceive length, area, angle, brightness the way they "should".
- Some illusions have been reclassified as systematic perceptual errors
  - e.g., brightness contrasts
    - (grey square on white background vs. on black background)
  - partly due to increase in our understanding of the relevant parts of the visual system
- Nevertheless, the visual system does some really unexpected things.

90

## Illusions of Linear Extent

- Mueller-Lyon (off by 25-30%)



- Horizontal-Vertical



91

## Illusions of Area

- Delboeuf Illusion



- Height of 4-story building overestimated by approximately 25%

92

## What are good guidelines for Infoviz?

- Use graphics appropriately
  - Don't use images gratuitously
  - Don't lie with graphics!
    - Link to original data
  - Don't conflate area with other information
    - E.g., use area in map to imply amount
- Make it interactive (feedback)
  - Brushing and linking
  - Multiple views
  - Overview + details
- Match mental models

93

## Tufte

- Principles of Graphical Excellence
  - Graphical excellence is
    - the well-designed presentation of interesting data – a matter of substance, of statistics, and of design
    - consists of complex ideas communicated with clarity, precision and efficiency
    - is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
    - requires telling the truth about the data.

– Edward Tufte, Visual Display of Quantitative Information, Graphics Press.  
 – Edward Tufte is a statistician and artist, and Professor Emeritus of Political Science, Statistics, and Computer Science at Yale University.

94

## Tufte's Notion of Data Ink Maximization

- What is the main idea?
  - draw viewers attention to the substance of the graphic
  - the role of redundancy
  - principles of editing and redesign
- What's wrong with this?
- What is he really getting at?

95

## Tufte Principle

Maximize the data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{data ink}}{\text{total ink used in graphic}}$$

Avoid "chart junk"

96

## Tufte Principles

- Use multifunctioning graphical elements
- Use small multiples
- Show mechanism, process, dynamics, and causality
- High data density
  - Number of items/area of graphic
  - This is controversial
    - White space thought to contribute to good visual design
    - Tufte's book itself has lots of white space

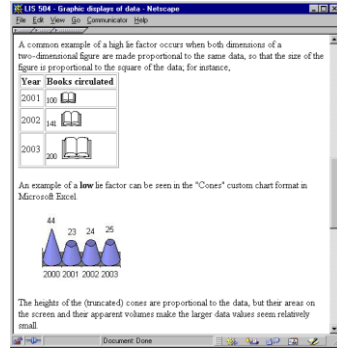
97



# Tufte's Graphical Integrity

- Some lapses intentional, some not
  - size of effect in graph
- Lie Factor =  $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$
- Misleading uses of area
- Misleading uses of perspective
- Leaving out important context
- Lack of taste and aesthetics

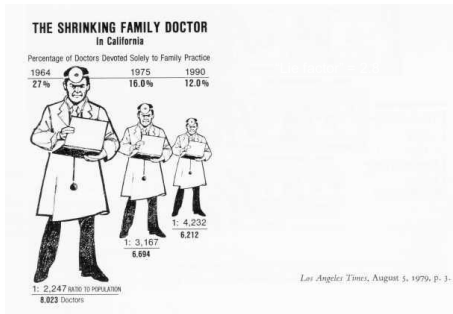
From Tim Craven's LIS 504 course  
[http://instruct.uwo.ca/fim-lis/504/504gra.htm#data-ink\\_ratio](http://instruct.uwo.ca/fim-lis/504/504gra.htm#data-ink_ratio)



98

99

## How to Exaggerate with Graphs (from Tufte '83)



100

## How to Exaggerate with Graphs (from Tufte '83)



101

## Animation

- “The quality or condition of being alive, active, spirited, or vigorous”
  - (dictionary.com)
- “A dynamic visual statement that evolves through movement or change in the display”
- “... creating the illusion of change by rapidly displaying a series of single frames”
  - (Roncarelli 1988).

102

## We Use Animation to...

- Tell stories / scenarios:
  - cartoons
- Illustrate dynamic process / simulation
- Create a character / an agent
- Navigate through virtual spaces
- Draw attention
- Delight

103

## Cartoon Animation Principles

- Chang & Unger '93
- Solidity (squash and stretch)
  - Solid drawing
  - Motion blur
  - Dissolves
- Exaggeration
  - Anticipation
  - Follow through
- Reinforcement
  - Slow in and slow out
  - Arcs
  - Follow through

104

## Why Cartoon-Style Animation?

- Cartoons' theatricality is powerful in communicating to the user.
- Cartoons can make UI engage the user into its world.
- The medium of cartoon animation is like that of graphic computers.

105

People are processing tools, too, especially when it comes to processing visual information

- In overview, good data visualization is arguably one of the most important challenges facing modern biologists.
- Although entire books have been written about data visualisation, it is worthwhile including an introduction to the production of visual representations of biological data.
  - This helps by aiding researchers in identifying patterns that relate to the underlying processes.
- Three simple but effective techniques, are:
  - Using HTML tables to list SWISS-PROT IDs.
  - Plotting an EMBL entry to show the arrangement of the Mer Operon genes.
  - Using Grace (a graph drawing program) to draw plots.
- All three of these techniques can be used on a home computer with minimal processing power and free software.

106

107

## Displaying Tabular Data Using HTML

- HTML can be used to generate meaningful, visual displays.
- There are two common situations in which HTML is used on the world wide web:
  - To create static web pages.
    - This is the simplest and most common use of HTML.
    - The browser requests the web page from the web server, which responds by sending an appropriately formatted text disk-file containing the HTML mark-up.
    - This disk-file is interpreted by the browser, producing a more visually pleasing representation of the document than flat text alone typically does.
  - To create on-the-fly dynamic web pages from server-side programs.
    - In this case, HTML web pages are generated for every request by a program executing on the web server.
    - Typically, input is provided to the program as part of the initial web request from the browser (using a HTML form).

108

## Displaying Tabular Data Using HTML

- A third situation involves producing HTML dynamically based on data parsed and processed from some other data source.
- Rather than storing the static HTML on a web server or producing the dynamic HTML on a web server, with this technique, HTML is saved to a local disk-file for later viewing as an "off-line" static web page.
- At some later date, the page can be made available through a web server if necessary.
- However, the use of HTML provides a local visualization representation of data.
- A custom program can take the data source and produces an HTML visualization as output.
- This allows for some very effective visualizations to be easily created.

109

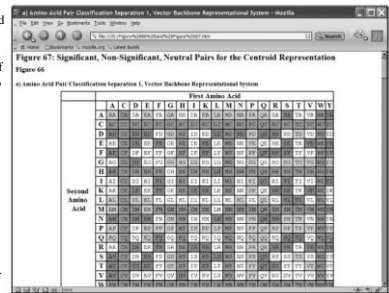
## Examples

- In the next two examples, The HTML used to present the visualisations was produced by custom Perl programs, which generated HTML disk-files.
- The disk-file were then viewed in the web browser.

### Example HTML visualisation: identifying amino acid states

Figure uses the background colour of the cell (which is grey-scaled in the figure because of printing restrictions) to identify which of four possible states each amino acid pair is in:

- Significantly different from an average state.
- Similar to an average state.
- Indeterminate.
- Those pairs with too few examples for an assessment to be made.



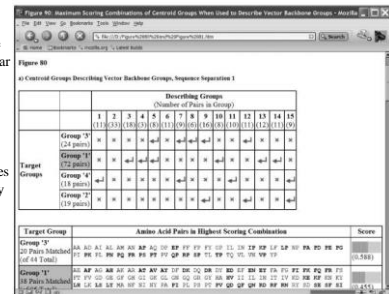
110

111

- The production of this HTML representation is the final high-level summary of a structural analysis pipeline, which itself contains four other stages.

### Example HTML visualisation: grouping amino acids

Figure describes which groups of amino acids from one analysis can be combined to give a similar group from another analysis. The bar graph display (used to represent the scores) is created from two small image disk-files that are included as many times as necessary using multiple HTML image tags (<IMG>). This is a simple "trick" that works very well.



112

113

- Despite the perceived complexity of these visual representations, the amount of HTML used in each is quite small:
  - each representation uses no more than ten individual HTML tags repeated over and over.

### Displaying SWISS-PROT identifiers

- In this visualisation, ID codes are sourced from the FASTA protein sequence disk-file containing the 55 Mer operon genes found in the SWISS-PROT database.
  - The idea is to extract and format these into a HTML table.
- This example was chosen for two reasons:
  - it is relatively straightforward and it demonstrates an important point, which is that producing the HTML mark-up is often the easy part.
  - What's harder is having the idea, the acquisition and extraction of the data, and its storage within a custom program, and so on.
- The custom program requires the FASTA disk-file to be in the format expected by theNCBI-BLAST package or supplied by the EBI SRS web-based service, for example:
  - sw/Q52109|MERA\_ACICA Mercuric reduct ...
  - MTTLKITGMTCDSCAAHVKEALEK ...

114

115

## Displaying SWISS-PROT identifiers

- Further, the program uses a combination of hard-coded HTML tags, as well as generated tags produced by the CGI module.
- Despite the availability of a table sub-routine with CGI, the use of hard-coded `<TABLE>` and `</TABLE>` is more convenient in certain situations.
- Another useful technique is the inclusion of newline characters as part of the resulting HTML disk-file.
- By default, web browsers generally ignore newline characters (as HTML provides the `<BR>` tag).
- Including them though makes the resultant HTML disk-file more readable by a human.

116

## Overview of the Mer Operon proteins in the SWISS-PROT database

| Gene | Accession Codes  | Gene IDs  |
|------|--|---|
| MERA | P00392, P30332, P08662, P08663, P16171, P17239, P30341, P84188, P84702, Q51772, Q52109, Q54465 | MERA_STRL1, MERA_BACCE, MERA_SHEFL, MERA_ACICA, MERA_THIFE, MERA_SEMA, MERA_ENTAG, MERA_PSEAF, MERA_STAEP, MERA_SHIFL, MERA_ALCSP, MERA_SHEFU |
| MERB | P08653, P08664, P16172, P30342, P71072, Q00792, Q05993, Q01082                                 | MERB_STAUP, MERB_SEMA, MERB_ECOLI, MERB_BACCE, MERB_PSEFU, MERB_STRL1, MERB_SHIFE, MERB_STAEP   |
| MERC | P04139, P04337, P22955   | MERC_THIFE, MERC_PSEAF, MERC_SHIFL  |
| MERD | P06689, P08654, P20102, P84703, Q51773, Q52110   | MERD_SALTI, MERD_PSEAF, MERD_SHIFL, MERD_SEMA, MERD_ACICA, MERD_PSEFL   |
| MERE | P06690   | MERE_PSEAF  |
| MERF | P04139, P04133, P13113, P04196, P84701, Q51770, Q52107, Q54463                                 | MERF_PSEFL, MERF_ACICA, MERF_SEMA, MERF_SALTI, MERF_SHIFE, MERF_PSEAF, MERF_ALCSP, MERF_SHEFU   |
| MERR | P06689, P07044, P13111, P22853, P22874, P22896, P30346   | MERR_SEMA, MERR_THIFE, MERR_STAUP, MERR_STRL1, MERR_BACCE, MERR_SALTI, MERR_PSEAF   |
| MERT | P04140, P04336, P08656, P13112, P30346, P04189, P84700, Q51769, Q52106, Q54462                 | MERT_PSEAF, MERT_ENTAG, MERT_STAUP, MERT_SEMA, MERT_ACICA, MERT_PSEFL, MERT_SALTI, MERT_STRL1, MERT_SHEFU, MERT_ALCSP                         |

117

## Creating High Quality Graphics With GD

- High-quality graphics are a very useful aid to data visualisation, both as an interactive, on-the-fly service attached to web pages, as well as when producing material for publications.
- One of the best interfaces to primitive graphic functions from within Perl programs is the **GD** module written by Lincoln D. Stein.
- This well-documented module hides a lot of the underlying complexity and links to the **gd graphics library** written by Tom Boutell.
- Depending on the functionality required, **gd** invokes a series of other libraries installed along with the operating system.
  - For example, to use **gd** to create PNG images requires the services of the **libpng** library (and the **zlib** library **libpng** calls).
  - Likewise, using **TypeType** fonts requires the installation of the **FreeType** library.

118

## Creating High Quality Graphics With GD

- For scientific work, **libpng** (and hence **zlib**) and **FreeType** are two of the most useful libraries to have installed.
- The **gd** library can produce also JPEG images if your system has **jpeglib** installed.
- However, as JPEGs tend to 'blur' images, the emphasis here is on producing PNG images, as they tend to preserve any crisp, sharp lines that exist within the image.

119

## Creating High Quality Graphics With GD

- Graphics primitives include:
  - circles, lines and rectangles, as well as colour definitions and direct pixel access techniques.
- These can be combined together to make more shapes or effects that are "less primitive" (more complex).

120

## Creating High Quality Graphics With GD

- Any missing libraries/functionality should be highlighted during the installation of GD.
- If something is missing, source it on the Internet and install it, before returning to the GD module and continuing the installation.
- Note that the installation of the GD module follows the standard Perl module installation process

```
$ perl Makefile.PL
$ make
$ make test
$ su
$ make install
$ <Ctrl-D>
```

121

## Creating High Quality Graphics With GD

- During the perl `Makefile.PL` step, the module asks if JPEG, FreeType and XPM support should be built.
- Be sure to answer "yes" to these questions so as to match the libraries that are installed.
- Included with the module is a demo directory.
- If the `ttf.pl` program within this directory executes with no errors, the module is very likely successfully installed.
- If it executes with errors, some additional installation work is still needed.
- Be sure to execute the `ttf.pl` program with this command-line:

```
$ ttf.pl | display
```

122

## The test image produced by the GD module

*Hello world!*  
*Hello world!*  
*Goodbye cruel world!*

123

## Using the GD module

- The GD module works around the concept of image canvases.
  - A **canvas** is a workspace upon which an image is manipulated.
  - Images can be created, loaded from an existing disk-file, drawn over, copied from other canvases or written to a disk-file in any of the supported graphic formats.
  - The module's documentation can be viewed on screen using the standard perl doc utility included with Perl

```
$ perldoc GD.pm
```

124

## Using the GD module - example

```
use GD;

my $image = new GD::Image( 100, 100 );
$white = $image->colorAllocate( 255, 255, 255 );
$black = $image->colorAllocate( 0, 0, 0 );
$red = $image->colorAllocate( 255, 0, 0 );
$blue = $image->colorAllocate( 0, 0, 255 );

$image->transparent( $white );
$image->interlaced( 'true' );
$image->rectangle( 0, 0, 99, 99, $black );
$image->arc( 50, 50, 95, 75, 0, 360, $blue );
$image->fill( 50, 50, $red );

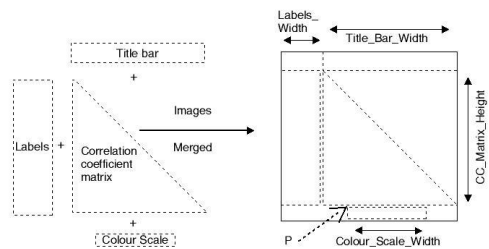
binmode STDOUT;

print $image->png;
```

125

## A sample image plan for a "heat map"

- Often, the most complex part of drawing images is working out where to put the various drawing components on the canvas, which is always the hardest part of drawing images using the graphics primitives.
- There are two techniques that can help reduce the problems caused by this complexity:
  - **Lots of planning**
    - take the time to plan a complex graphic in a vector graphics program or on paper before starting to code.
    - Make a note of some meaningful variable names and annotate the drawing with the values of any constant "off-sets".
  - **Use multiple canvases**
    - the GD module can create a series of canvases, each of which can be manipulated separately.
    - These are then merged together into one parent canvas prior to producing the image on STDOUT.
    - This prevents many "off-sets" in each graphics call and moves the problem to one "copy image" statement for each sub-canvas.
      - These calls can still be formidable, but it is one that has to be right only once.
      - This is where the image plan described in the previous point is most helpful.



126

127

- An important point to consider is the eventual size of the image.
- A general rule of thumb suggests the bigger the image, the better.
- It is straightforward to convert a large image to a smaller size by averaging the information already present than it is to expand an image.
- The "Cost of Canvas" in the GD module is low, allowing large images to be generated.

Producing plans avoids problems before problems surface

128

129

## Displaying genes in EMBL entries

- The program is designed to generate the base graphic for the image from the data contained in the original EMBL entry as a guide.
- The program is designed to demonstrate the GD module's drawing capabilities in its most general form.
  - However, the program is not a general solution that can be used with any EMBL entry.
- The code uses the object-orientated interface provided by the GD module.
- Although strange at first, the object-oriented syntax is comprehensible and easy to relate to and use.
- The generated image is large.
  - The image is over 8000 pixels wide and is designed to be resampled (discussed below) rather than used as-is.
- The program uses a specific font, identified in the code as `albr85w.ttf` within the `/windows/C/WINDOWS/Fonts` directory.
  - This font may or may not be installed on every computer and an alternative may need to be substituted.
- Note that the GD module provides a generic font called `Generic.ttf`.

## A plot of the interesting genes identified in EMBL entry ISTN501



130

131

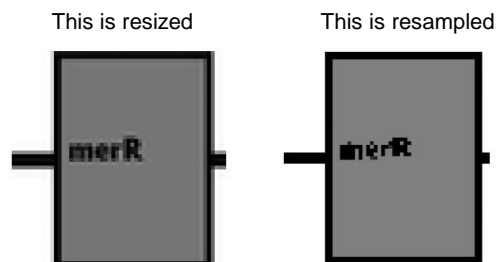
## Introducing mogrify

The `mogrify` utility, among other things, can transform images by reducing the number of pixels they use, when invoked with either the `-size <geometry>` or `-sample <geometry>` switches.

```
http://www.imagemagick.org/

$ mogrify -resize 1600 Embl_sequence_graphic.png
$ mogrify -resize x100 Embl_sequence_graphic.png
$ man mogrify
$ cp Embl_sequence_graphic.png
  Embl_sequence_graphic.original.png
$ mogrify -resize 1600 Embl_sequence_graphic.png
```

## The difference between resampling and resizing.



132

133

## Plotting Graphs

- Plotting data in the form of a pictorial graph is often a very useful way to view numerical data.
  - Technical Commentary: We use the word "pictorial" here as the word "graph" is also used in the context of graph theory.
  - This is a general mathematical description of interlinked vertices (points or nodes) that are connected by edges (links).
  - This terminology is used from time to time in Bioinformatics and bioscience literature to represent networks of how things are related to each other or linked together in pathways.
  - Not knowing the difference between the two can be confusing.

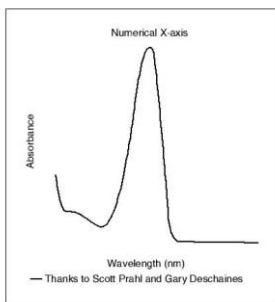
134

## Graph plotting using the GD::Graph modules

- The `GD::Graph` and `GD::Graph3d` modules, available on CPAN, interface with the GD library, providing a convenient, programmer-controlled way to produce high-quality graphs.
- The `GD::Graph` module produces the standard set of graph types:
  - bars, stacked bars, lines, XY points and pies.
- The `GD::Graph3d` module adds extra shading and perspective, simulating a 3D look and feel.

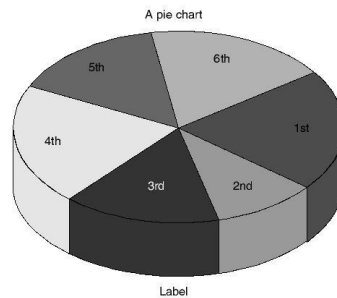
135

### Example line graph from the GD::Graph module



136

### Example pie chart from the GD::Graph module



137

## Graph plotting using Grace

- Grace is designed to be used for scientific graphing, and is especially good at XY scatter plots.
- It is freely available from the following web-site: <http://plasma-gate.weizmann.ac.il/Grace/>
- The Grace executable is called `xmgrace`.
- There are three common approaches to plotting graphs with Grace:
  - Interactive graph generation using a GUI-based application program.
  - Command-line "batch" plotting using the `xmgrace` command-line utility.
  - Programmer-generated graphs using the `Chart::Graph::Xmgrace` module from CPAN.

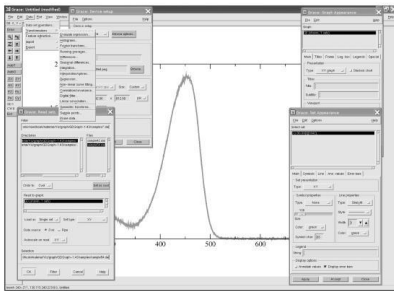
138

## Interactive plotting

- Next Figure shows the "Absorbance" test dataset within the GUI-based Grace program.
- To recreate this graph, load the required data, as follows:
  - From the menu, select **Data**, then **Import**, then **ASCII** to identify the `sample54.dat` disk-file included with the `GD::Graph` module.
  - Highlight the disk-file name in the `Grace:Read Sets` dialogue box (as shown in the screen shot), then click the **OK** button.
  - The data set loads and the graph appears, autoscaled to fit within the display.
  - Double click on the drawn line to open the `Grace:Set Appearance` dialoguebox.
  - Set the **Line Width**, in the **Line properties** section on the **Main** panel, to 3.
  - Sit back and admire your handiwork!

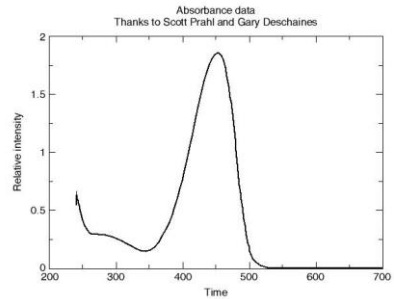
139

### The GUI-based Grace application program



140

### The "Absorbance" image as produced by Chart::Graph::Xmgrace

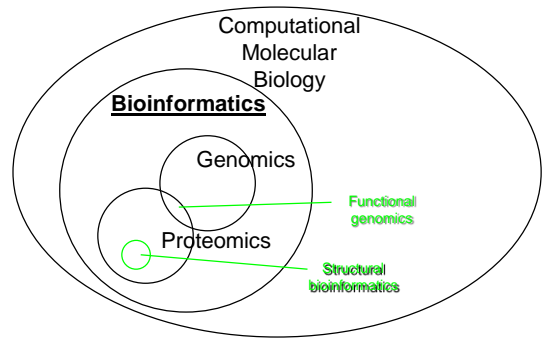


141

### Where To From Here

- Following slides illustrate visualization examples in bioinformatics

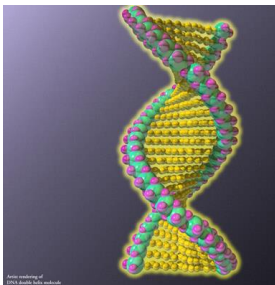
### Bioinformatics Nomenclature



142

143

### DNA is the blueprint for life



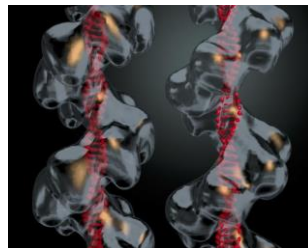
- Every cell in your body has 23 *chromosomes* in the nucleus
- The *genes* in these chromosomes determine all of your physical attributes.

Image source: Crane digital, <http://www.cranedigital.com/>

144

### Mapping the Genome

- The human genome project has provided us with a draft of the *entire human genome*.



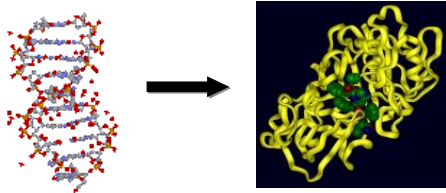
- Four bases: A, T, C, G
- 3.12 billion base-pairs
- 99% of these are the same
- *Polymorphisms* = where they differ

145



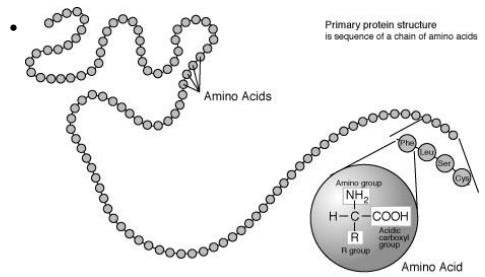
## From sequence to structure

- Genes contain the protocols for construction of proteins:



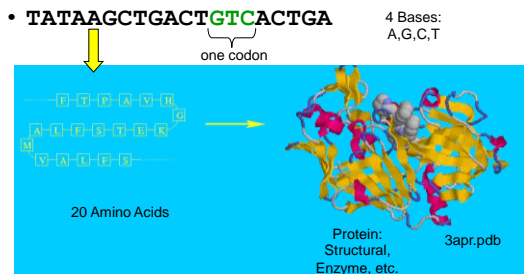
146

## Proteins are chains of amino acids



147

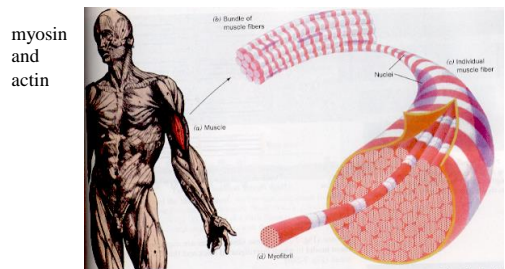
## Genomic information: from genes to proteins



148

## Proteins: Molecular machinery

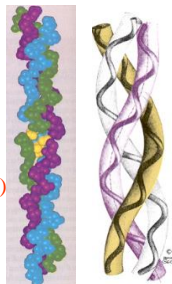
- Proteins in your muscles allows you to move:



149

## Proteins: Molecular machinery

- Enzymes (digestion, catalysis)
- Structure (collagen)
- Immune response (antibodies)
- Self-recognition (MHC)



150

## Proteins: Molecular machinery

- Signaling (hormones, kinases)
- Transport (energy, oxygen)

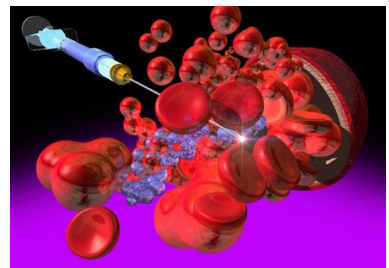
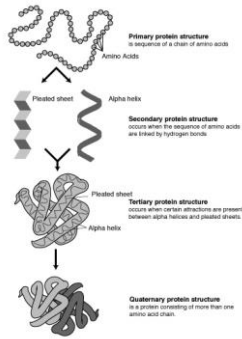


Image source: Crane digital, <http://www.cranedigital.com/>

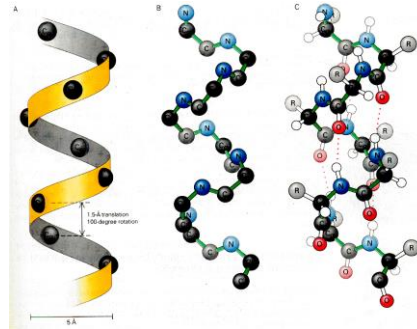
151

## Shape is paramount to function



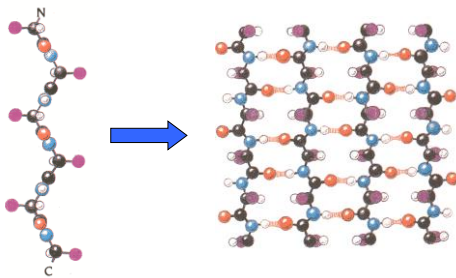
152

## The alpha helix



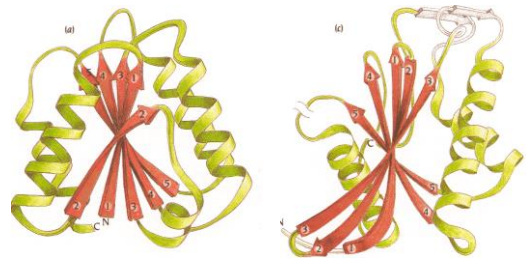
153

## The beta strand (& sheet)



154

## Protein Domains



155

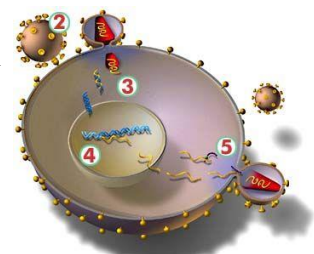
## Why are we interested in protein structure?

- Misfunctioning or malformed proteins cause disease
- Proteins serve as **drug targets**
  - Bacterial pathogens have distinct proteins that they need to survive and proliferate
  - Viruses hijack our molecular machinery to construct proteins using their own DNA or RNA messages

156

## Example Case: HIV Protease

1. Exposure & infection
2. HIV enters your cell
3. Your own cell reads the HIV “code” and creates the HIV proteins.
4. New viral proteins prepare HIV for infection of other cells.

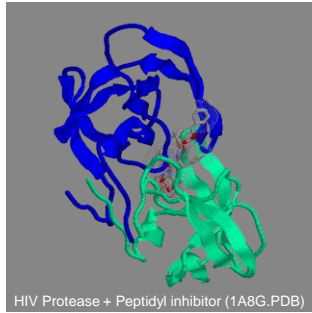


© George Eade, Eade Creative Services, Inc.  
<http://whyfiles.org/035aids/index.html>

157

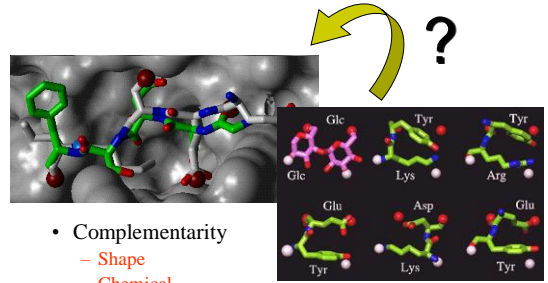
## HIV Protease as a drug target

- Many drugs bind to protein active sites.
- This HIV protease can *no longer* prepare HIV proteins for infection, because *an inhibitor is already bound* in its active site.



158

## Drug Lead Screening & Docking



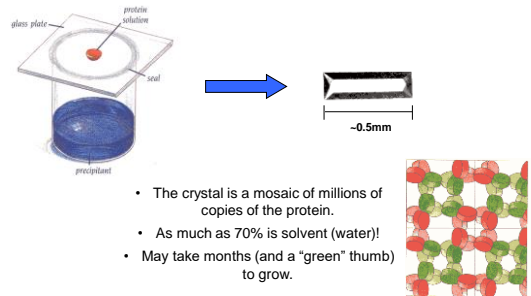
- Complementarity
  - Shape
  - Chemical
  - Electrostatic

159

## Importance of Molecular Graphics

- The only window we have into the world of protein structure
- Proteins are too small to view in molecular detail using any modern microscopy techniques
- Data is collected with two techniques
  - X-Ray crystallography
  - Nuclear Magnetic Resonance (NMR)

## X-Ray Crystallography

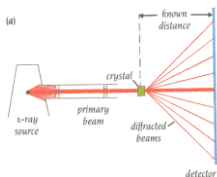


- The crystal is a mosaic of millions of copies of the protein.
- As much as 70% is solvent (water)!
- May take months (and a "green" thumb) to grow.

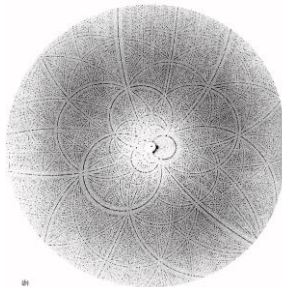
160

161

## X-Ray diffraction



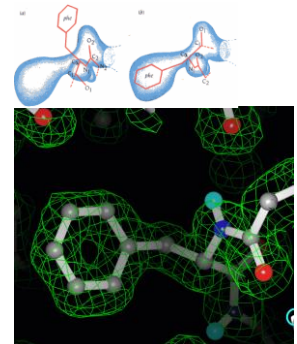
- Image is averaged over:
  - Space (many copies)
  - Time (of the experiment)



162

## Electron Density Maps

- Resolution is dependent on the quality/regularity of the crystal
- R-factor is a measure of "leftover" electron density
- Solvent fitting
- Refinement



163

# The Protein Data Bank

# A PDB structure:

- <http://www.rcsb.org/pdb/>

|      |    |     |     |   |   |        |        |         |      |       |          |
|------|----|-----|-----|---|---|--------|--------|---------|------|-------|----------|
| ATOM | 1  | N   | ALA | E | 1 | 22.382 | 47.782 | 112.975 | 1.00 | 24.09 | 3APR 213 |
| ATOM | 2  | CA  | ALA | E | 1 | 22.957 | 47.648 | 111.613 | 1.00 | 22.40 | 3APR 214 |
| ATOM | 3  | C   | ALA | E | 1 | 23.572 | 46.251 | 111.545 | 1.00 | 21.32 | 3APR 215 |
| ATOM | 4  | O   | ALA | E | 1 | 23.948 | 45.688 | 112.603 | 1.00 | 21.54 | 3APR 216 |
| ATOM | 5  | CB  | ALA | E | 1 | 23.932 | 48.787 | 111.380 | 1.00 | 22.79 | 3APR 217 |
| ATOM | 6  | N   | GLY | E | 2 | 23.656 | 45.723 | 110.336 | 1.00 | 19.17 | 3APR 218 |
| ATOM | 7  | CA  | GLY | E | 2 | 24.216 | 44.393 | 110.087 | 1.00 | 17.35 | 3APR 219 |
| ATOM | 8  | C   | GLY | E | 2 | 25.653 | 44.308 | 110.579 | 1.00 | 16.49 | 3APR 220 |
| ATOM | 9  | O   | GLY | E | 2 | 26.258 | 45.296 | 110.994 | 1.00 | 15.35 | 3APR 221 |
| ATOM | 10 | N   | VAL | E | 3 | 26.213 | 43.110 | 110.521 | 1.00 | 16.21 | 3APR 222 |
| ATOM | 11 | CA  | VAL | E | 3 | 27.594 | 42.879 | 110.975 | 1.00 | 16.02 | 3APR 223 |
| ATOM | 12 | C   | VAL | E | 3 | 28.569 | 43.613 | 110.055 | 1.00 | 15.69 | 3APR 224 |
| ATOM | 13 | O   | VAL | E | 3 | 28.429 | 43.444 | 108.822 | 1.00 | 16.43 | 3APR 225 |
| ATOM | 14 | CB  | VAL | E | 3 | 27.834 | 41.363 | 110.979 | 1.00 | 16.66 | 3APR 226 |
| ATOM | 15 | CG1 | VAL | E | 3 | 29.259 | 41.013 | 111.404 | 1.00 | 17.35 | 3APR 227 |
| ATOM | 16 | CG2 | VAL | E | 3 | 26.811 | 40.649 | 111.850 | 1.00 | 17.03 | 3APR 228 |



164

165

## NMR structures

- Uses magnetic resonance to determine a set of constraints on the inter-atomic distances.
- Multiple models can satisfy these constraints:



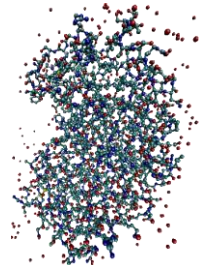
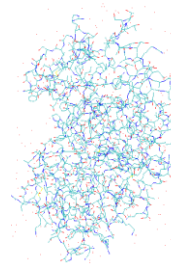
166

167

## Views of a protein

Wireframe

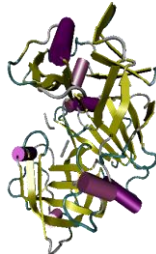
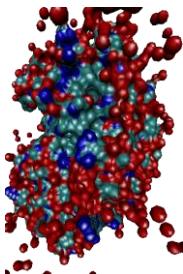
Ball and stick



## Views of a protein

Spacefill

Cartoon



CPK colors  
 Carbon = green, black, or grey  
 Nitrogen = blue  
 Oxygen = red  
 Sulfur = yellow  
 Hydrogen = white

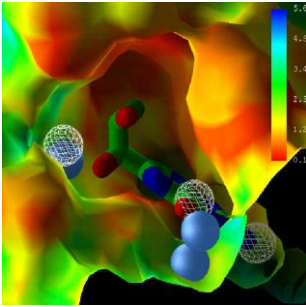
168

169

## Two primary directions

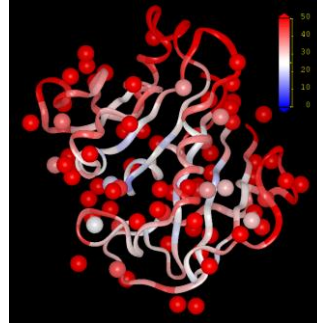
- Exploration
  - Inspire new computational approaches
  - Verify results of computational methods
- Expression
  - Demonstrate results to a scientific audience

## Color Mapping



170

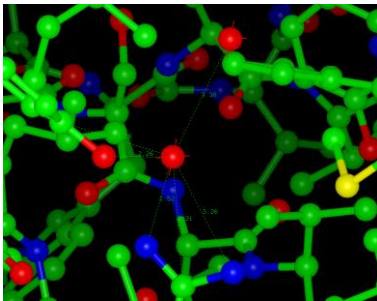
## Temperature Factor (BVAL)



The backbone of dihydrofolate reductase (IDR2) is shown as ribbons colored according to crystallographic temperature factor (B-value).

171

## Atomic Density

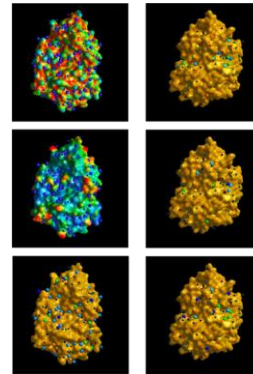


A water molecule in the ligand-free structure of dihydrofolate reductase (IDR2).

The atomic density of this water molecule is 5.

172

## Exploratory data analysis



173

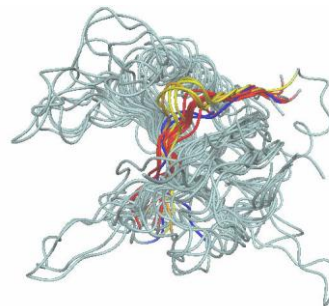
## Molecular Modeling

- Determining the structure of a protein from only the sequence of amino acids
- *Protein folding* –  
– a grand challenge problem
- *Comparative modeling* –  
– find other proteins of similar sequence and look at how they fold

```
MAVIPKARVLGFIAVGDGDHCTANQGPLC
MAVIPKARIAAGFIAV-IGEDHCTANQGPLC
MA--PKARVLGFILIGDGDGHCTANQGPLC
MAVIPKARVLGFIVVGGDD--TANQGPLC
```

174

## Comparative modeling

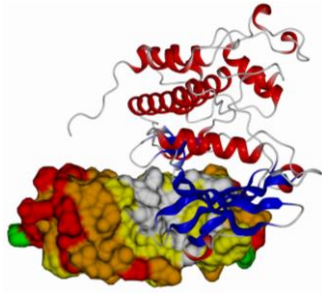


Clustered OB fold fragments reveal three distinct fragment clusters for modeling

175

## Conservation surface mapping

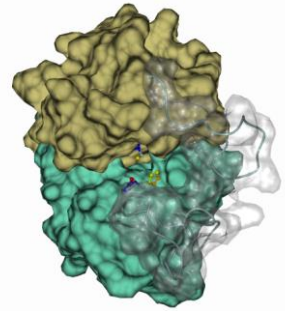
Cyclin-dependant protein kinase (CDK-6, ribbons) & inhibitor, colored according to evolutionary conservation.



176

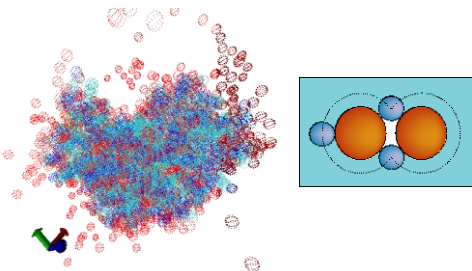
## Computational docking

Predicted binding mode for cathepsin B inhibition by phenformin



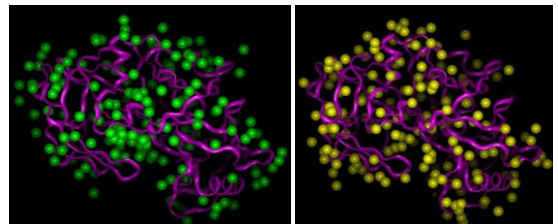
177

## Dot surfaces and probe points



178

## Probe Site Generation

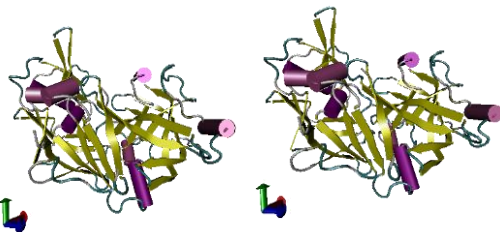


Aspartic protease (2apr) with crystallographically observed and computer-generated water molecules.

179

## 3D Visualization

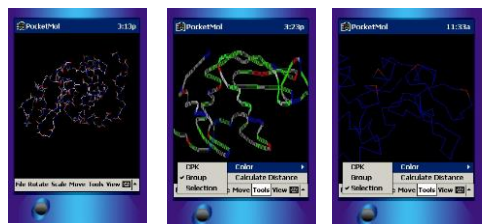
The state of the art:



180

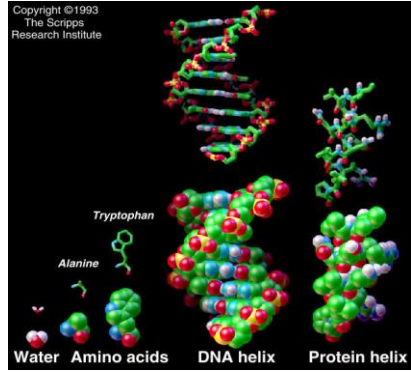
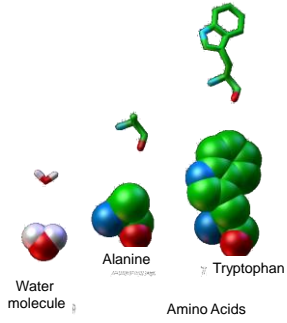
## PocketMol

- 3D Visualization for PocketPC



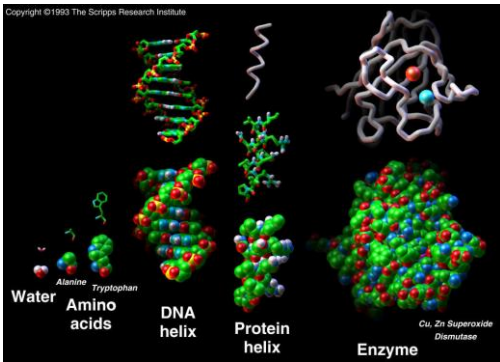
181

# Expression:

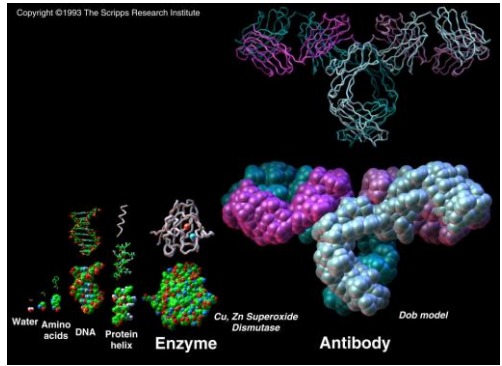


182

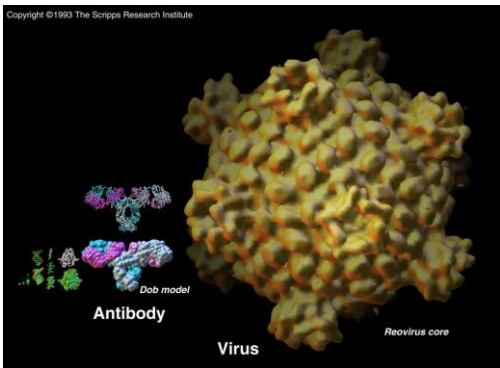
183



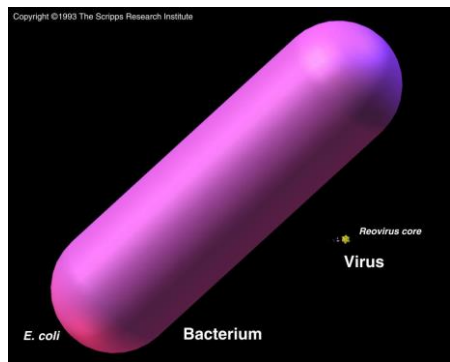
184



185



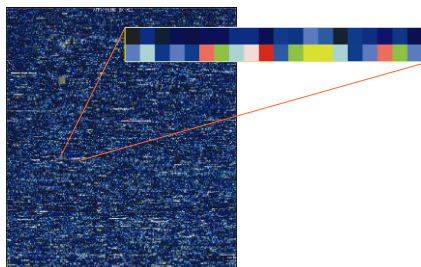
186



187

## Other major applications

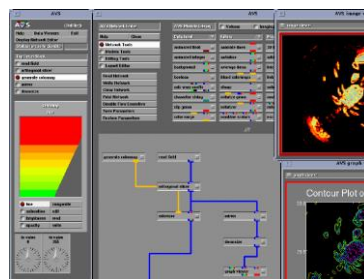
- Microarray (*Genechip*) expression data visualization



188

## Molecular Visualization Tools

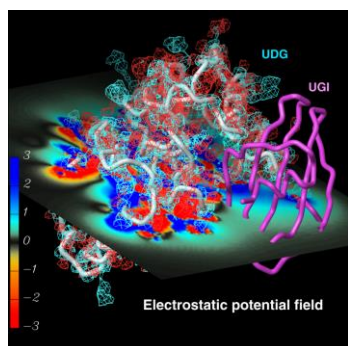
- Rasmol
- VMD
- Dino
- MSI/Sybyl
- AVS5 & AVS/Express Visualization Edition



189

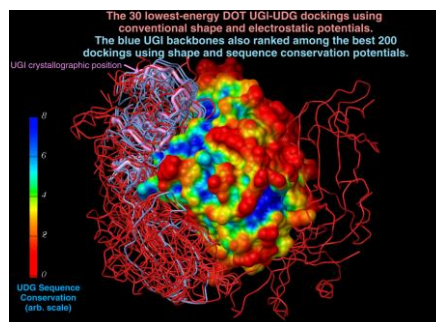
## The future of molecular visualization

- High-dimensionality data:



190

## Rich Information Content



191