Introduction to Protein Structure

- 1-D world of nucleotide structure and amino acid sequences

➔ now enter to ➔

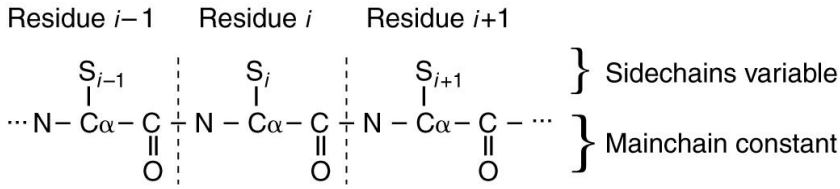
- 3-D world of molecular structures

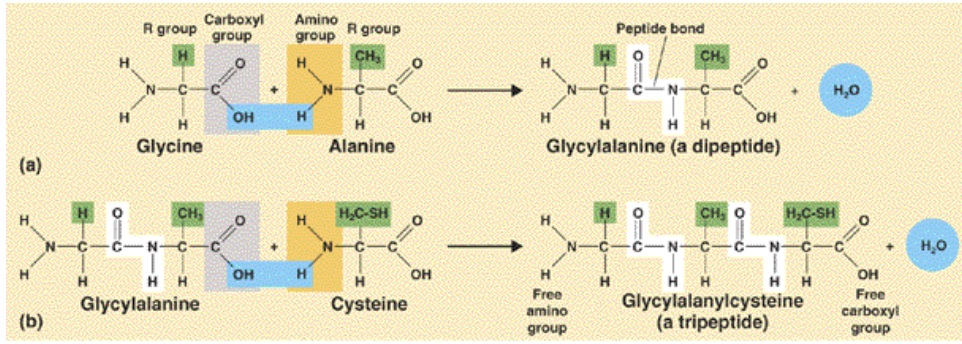
Proteins play a variety of roles in life process

- Structural proteins
- Enzymes: proteins that catalyze (催化) chemical reactions
- Transport and storage proteins
- Regulatory proteins
- Proteins that control gene transcription
- Proteins that involved in recognition, including cell adhesion (黏著) molecules,
- Antibodies and other protein of the immune system

- Proteins are ***large molecules***.
- In many cases only a small part of the structure – an ***active site*** – is directly functional, the rest existing primarily to create and fix the spatial relationship among the active site residues.
- Proteins evolve by ***structural changes***, produced by ***mutations*** in the amino acid sequence and ***genetic rearrangements***, that bring together different combinations of structural subunits.

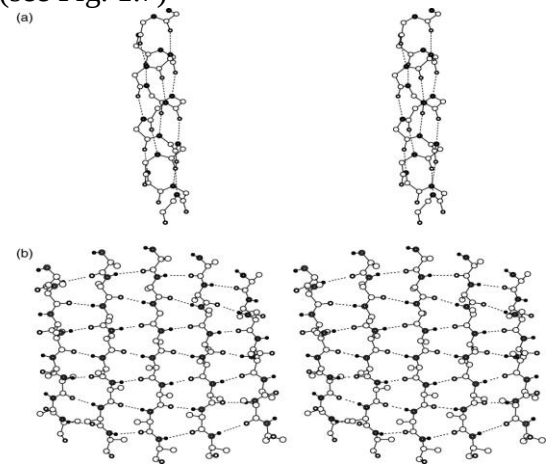
- ~ ***85,000*** protein structures are now known
- Most were determined by ***X-ray crystallography*** or ***NMR*** (nuclear magnetic resonance)
- Few were determined by electron microscopy and others
- Chemically, protein molecules are long polymers typically containing several thousand atoms, composed of a uniform repetitive ***backbone*** (or ***mainchain***) with a particular ***sidechain*** attached to each residue (see Fig. 1.6)
- Amino acid sequence of a protein records the succession of sidechains.



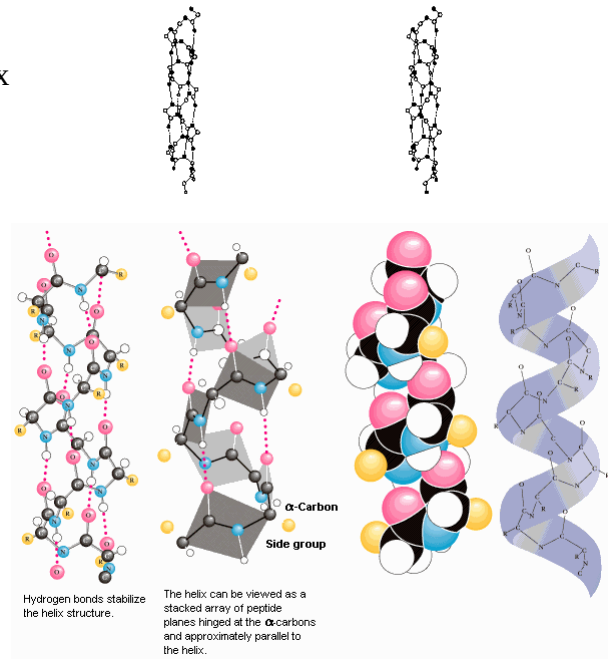


- The polypeptide chain folds into a curve in space
- The course of the chain defining a ***folding pattern***
- A great variety of folding patterns: a number of common structural features
- α helices and β sheets (see Fig. 1.7)
  - 螺旋 (helix)
  - 摺板 (sheet)

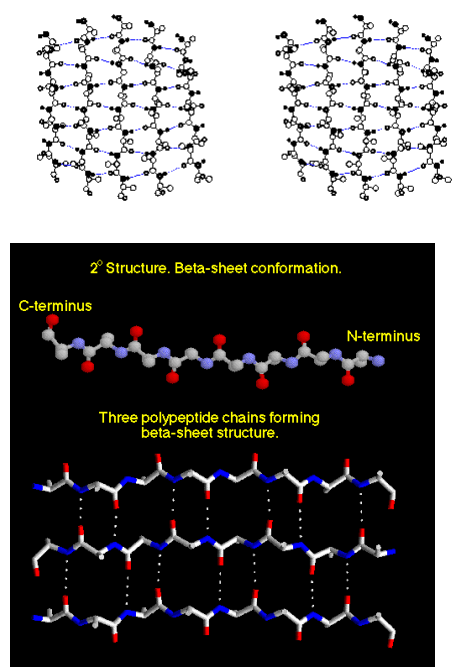
• Folding may be thought of as a kind of intramolecular condensation or crystallization



α helix



β sheet

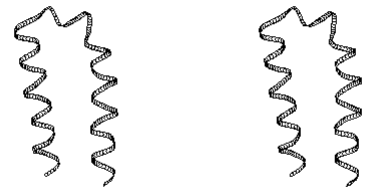


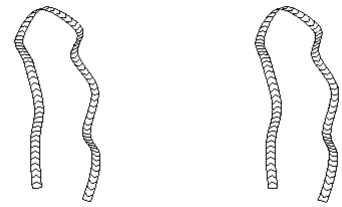
Hierarchical nature of protein architecture

- ***Primary structure***: the amino acid sequence – the set of primary chemical bonds
- ***Secondary structure***: the assignment of helices and sheets – the hydrogen-bonding pattern of the mainchain
- ***Tertiary structure***: the assembly and interactions of the helices and sheets
- ***Quaternary structure***: for proteins composed of more than one subunit, the assembly of the monomers (單體)

Additional levels to the hierarchy

- ***Supersecondary structures***: include the alpha-helix hairpin, the beta-hairpin, and the beta-alpha-beta unit. (Fig. 1.8)
- ***Domains***: many proteins contain compact units within the folding pattern of a single chain, that look as if they should have independent stability. (Fig. 1.9)
- ***Modular proteins*:** are multidomain proteins which often contain many copies of closely related domains.
  - Domain recur in many proteins in different structural contexts; that is, different modular proteins can ‘mix and match’ sets of domains.





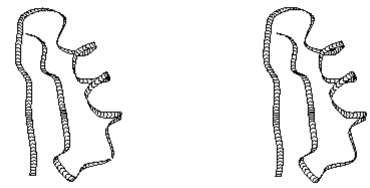
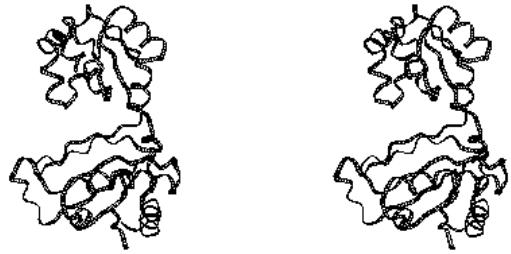
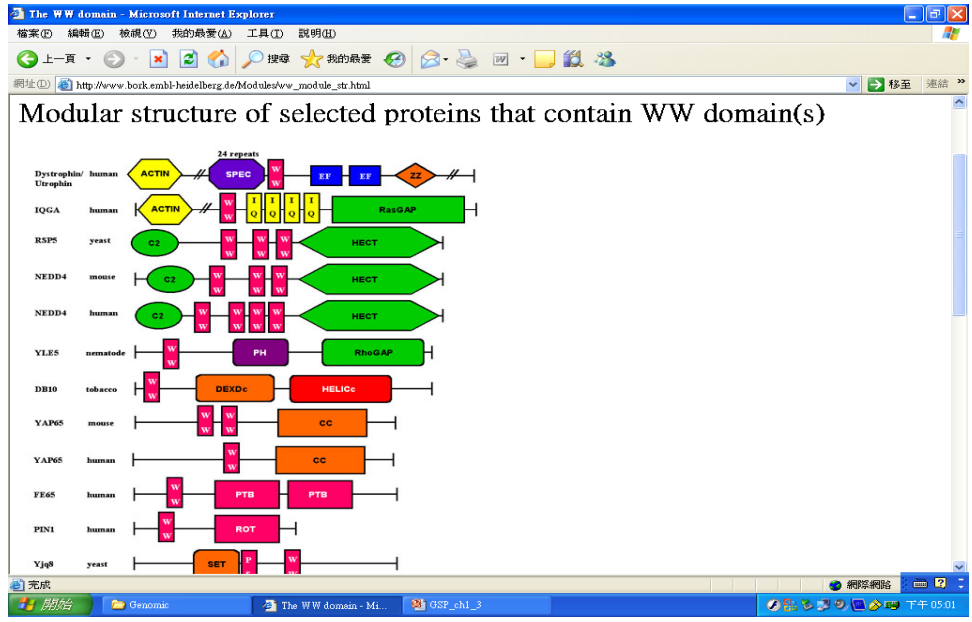


Fig. 1.9 RNA binding protein L1:



Multidomain proteins



Classification of protein structures

- The most general classification of families of protein structures is based on the ***secondary and tertiary*** structures
- Classification of protein structures occupies a key position in bioinformatics, not least as a bridge between sequence and function.

Class	Characteristic
α-helical	secondary structure exclusively or almost exclusively α-helical
β-sheet	secondary structure exclusively or almost exclusively β-sheet
$\alpha + \beta$	α-helices and β-sheets separated in different parts of the molecule; absence of β-α-β supersecondary structure
α/β	helices and sheets assembled from β-α-β units
α/β-linear	line through centres of strands of sheet roughly linear
α/β-barrels	line through centres of strands of sheet roughly circular
little or no secondary structure	

Fig 1-10a: engrailed homeodomain [1enh]:

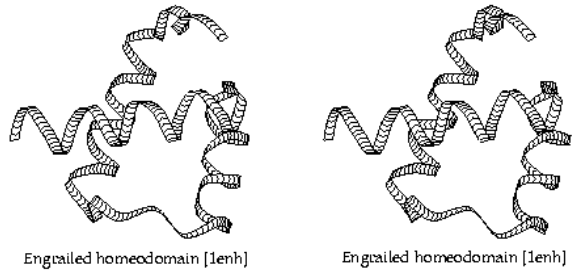


Fig 1-10b: second calponin homology domain from utrophin [1bhd]:

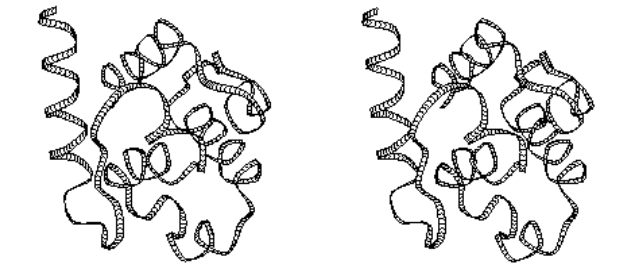
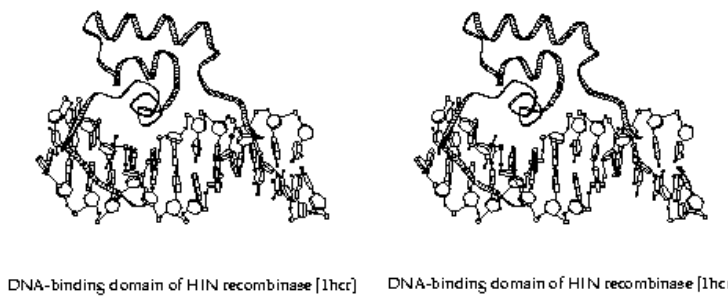
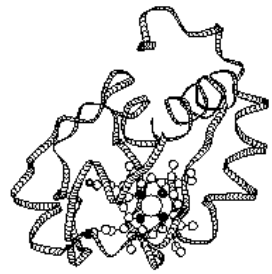


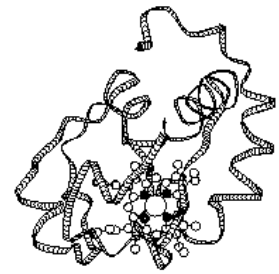
Fig 1-10c: HIN recombinase, DNA-binding domain [1hcr]:



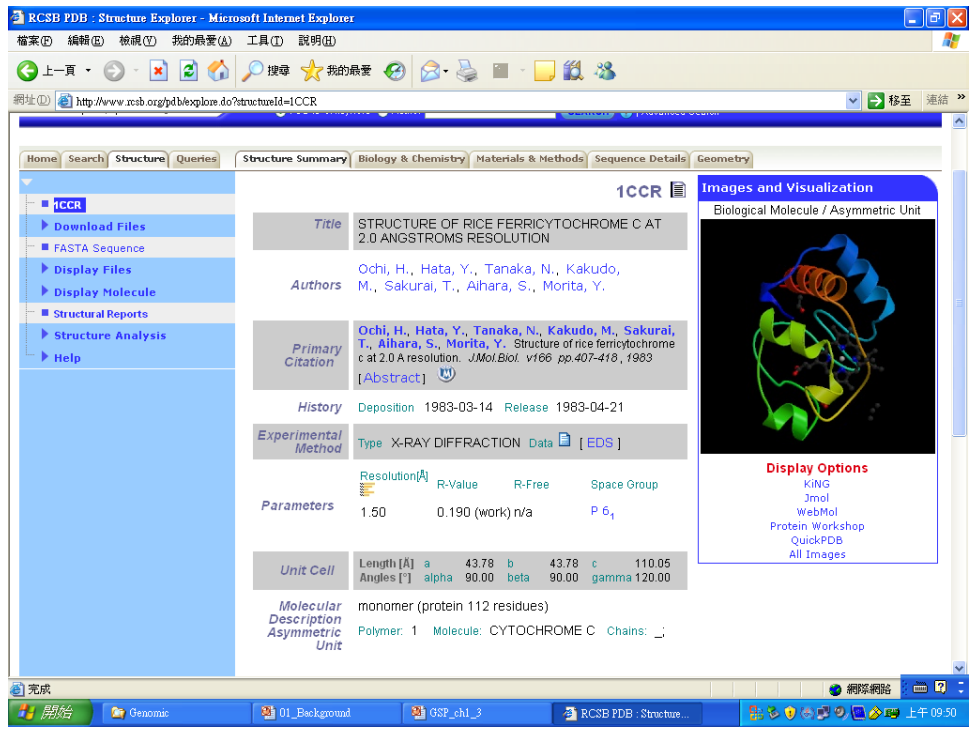
(d) Rice embryo cytochrome *c* [1ccr]

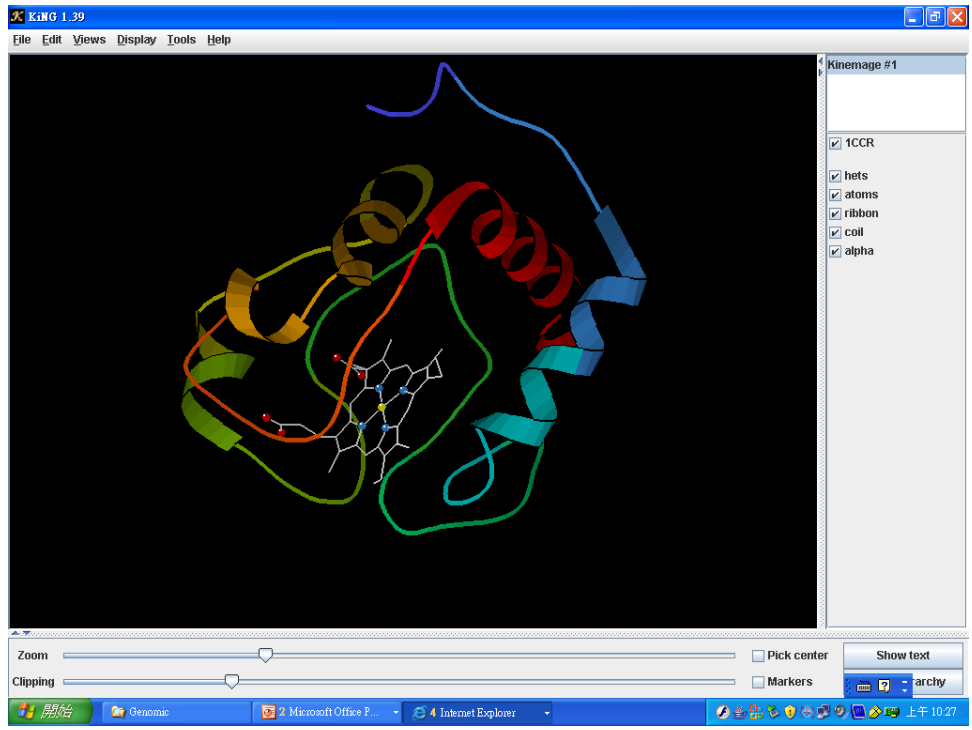


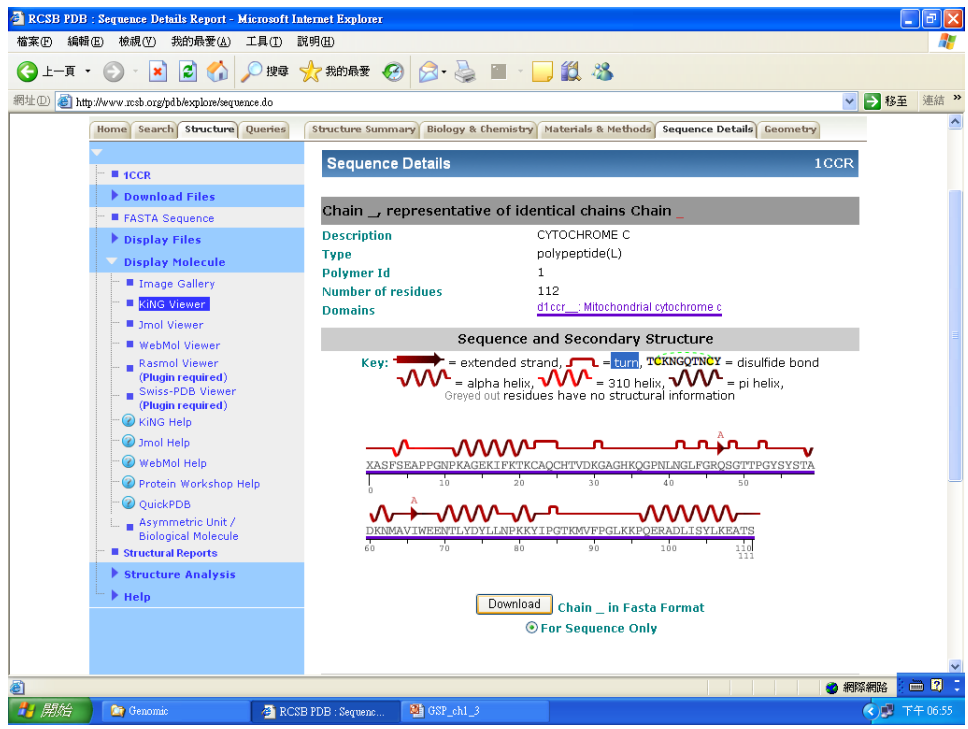
Cytochrome c, rice [1ccr]

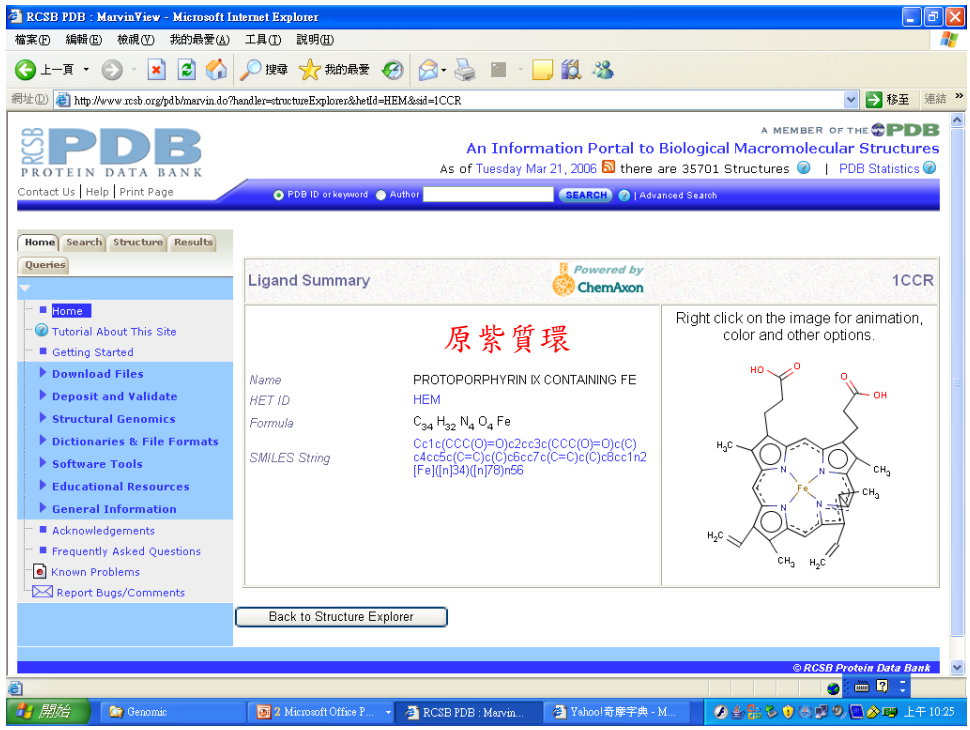


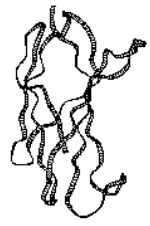
Cytochrome c, rice [1ccr]




KiNG 1.39
File Edit Views Display Tools Help
Kinemage #1
1CCR
hets
atoms
ribbon
coil
alpha
Zoom
Clipping
Pick center
Markers
Show text

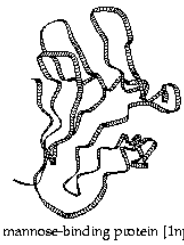

RCSB PDB : Sequence Details Report - Microsoft Internet Explorer
http://www.rcsb.org/pdb/explore/sequence.do
Home Search Structure Queries Structure Summary Biology & Chemistry Materials & Methods Sequence Details Geometry
1CCR
Download Files
FASTA Sequence
Display Files
Display Molecule
Image Gallery
KiNG Viewer
Jmol Viewer
WebMol Viewer
Rasmol Viewer (Plugin required)
Swiss-PDB Viewer (Plugin required)
KiNG Help
Jmol Help
WebMol Help
Protein Workshop Help
QuickPDB
Asymmetric Unit / Biological Molecule
Structural Reports
Structure Analysis
Help
Sequence Details 1CCR
Chain _, representative of identical chains Chain _
Description CYTOCHROME C
Type polypeptide(L)
Polymer Id 1
Number of residues 112
Domains d1ccr__: Mitochondrial cytochrome c
Sequence and Secondary Structure
Key: = extended strand, = turn, TCKNGQTNCY = disulfide bond
= alpha helix, = 310 helix, = pi helix,
Greyed out residues have no structural information
XASFSEAPPGNPKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTPGYSYSTA
DKNMAVIWEENTLYDYLLNPKKYIPGTKMVFPGLKKPQERADLISYLKEATS
Download Chain _ in Fasta Format
For Sequence Only





Fibronectin III domain [1fna]

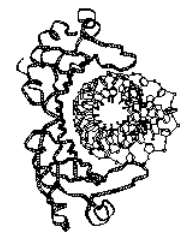
Fibronectin III domain [1fna]



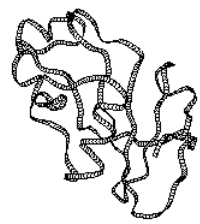
mannose-binding protein [1npl]

mannose-binding protein [1npl]

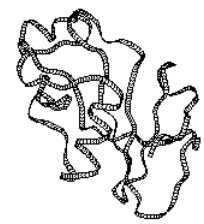
TATA-box-binding protein [1cdw]



TATA-box-binding protein [1cdw]



barnase [1brn]



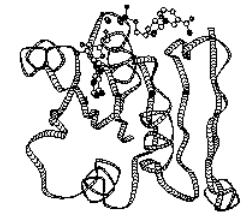
barnase [1brn]

OB-domain from Lys-tRNA synthetase [1bbw]

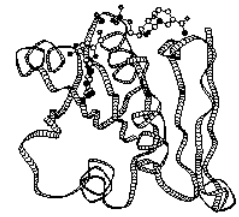
OB-domain from Lys-tRNA synthetase [1bbw]

Scytalone dehydratase [3std]

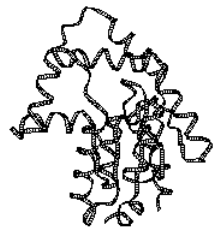
Scytalone dehydratase [3std]



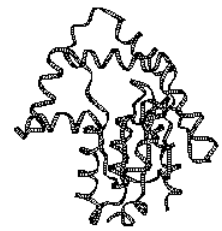
alcohol dehydrogenase, NAD-binding domain [1ee2]



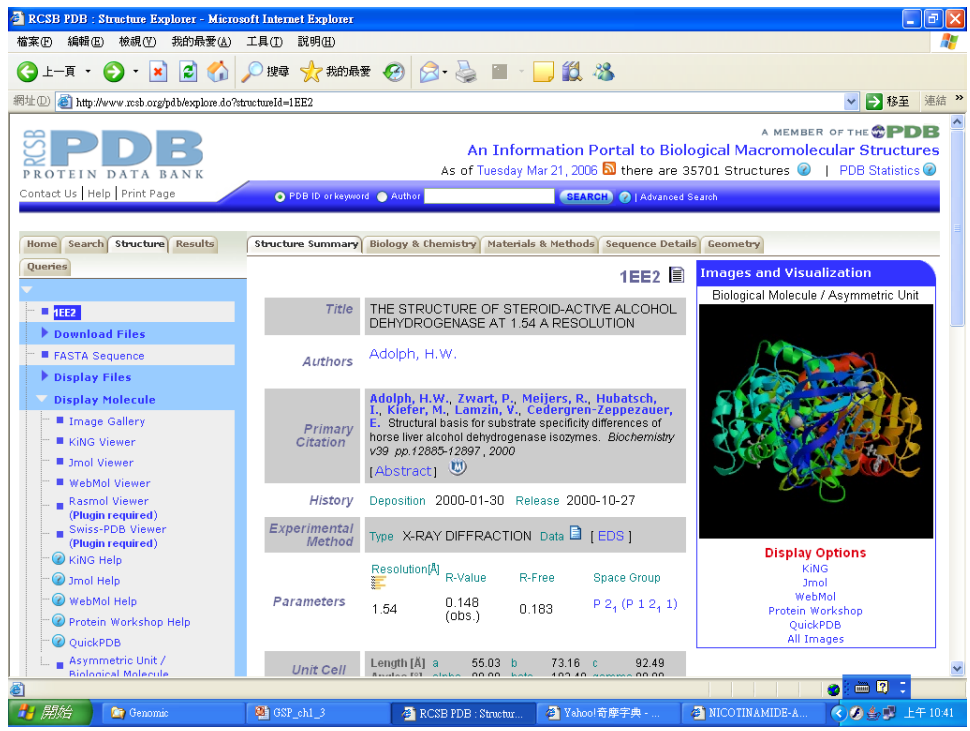
alcohol dehydrogenase, NAD-binding domain [1ee2]

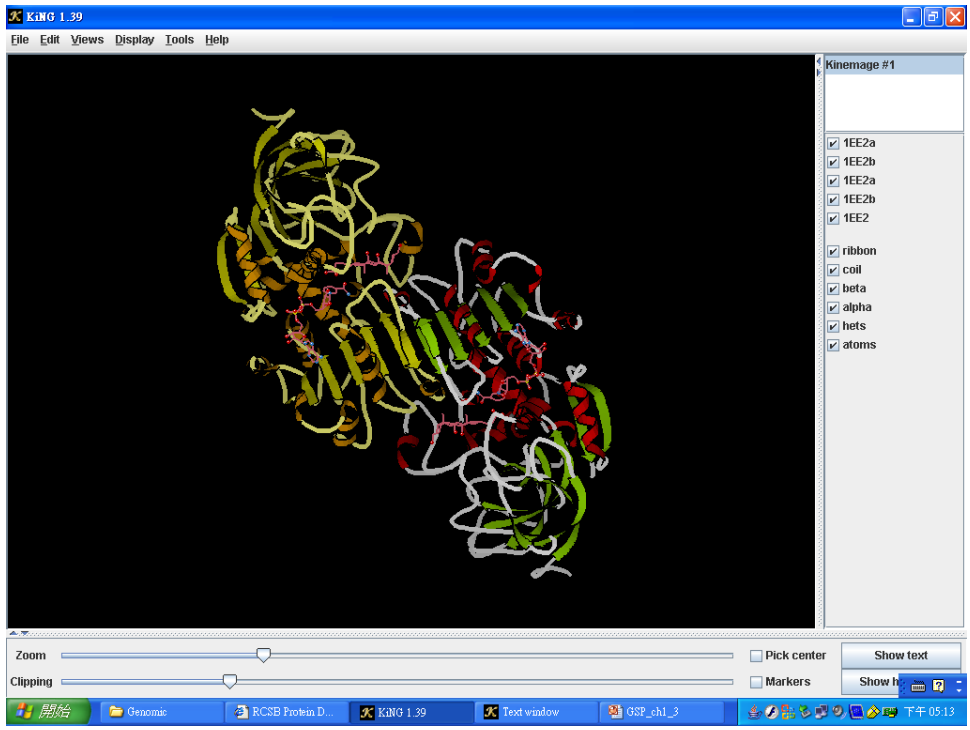


Adenylate kinase [3adk]



Adenylate kinase [3adk]

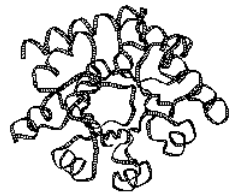




輔一菸鹼胺腺吟雙核酸(Nicotinamide adenine dinucleotide, NAD)

Chemotaxis receptor methyltransferase [1af7]

Chemotaxis receptor methyltransferase [1af7]



Thiamine phosphate synthase [2tps]

Thiamine phosphate synthase [2tps]

pancreatic spasmolytic polypeptide [2psp]

pancreatic spasmolytic polypeptide [2psp]

Web resources

- The Worldwide Protein Data Bank (wwPDB) http://www.wwpdb.org/
- The Research Collaboratory for Structural Bioinformatics (RCSB) (USA) http://www.rcsb.org/
- The Macromolecular Structure Database (MSD) (UK) http://www.ebi.ac.uk/pdbe/
- The protein databank Japan http://www.pdbj.org/
- BMRB (USA) http://www.bmrb.wisc.edu/
- Structural Classification of Proteins (SCOP) http://scop.mrc-lmb.cam.ac.uk/scop/
- The Molecular Modeling DataBase (MMDB) http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

Protein structure prediction and engineering

- Amino acid sequence of a protein dictates its 3D structure
- If amino acid sequences contain sufficient information to specify 3D structures of proteins, it should be possible to ***devise an algorithm to predict protein structure from amino acid sequence.***
  - This has proved elusive (難以理解的).

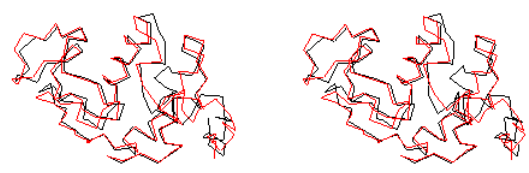
Less-ambitious goals:

- **Secondary structure prediction** – which segments of the sequence form helices and which form strands of sheet?
- **Fold recognition** – Given a library of known protein structures and their amino acids sequences, and the amino acid sequence of a protein of unknown structure, can we find the structure in the library that is most likely to have a folding pattern similar to that of the protein of unknown structure?
- **Homology modelling** – If the sequences of two homolgous proteins have 50% or more identical residues in an optimal alignment, the structures are likely to have similar conformations over more than 90% of the model.

Aligned sequences and superposed structures of two related proteins: Alignment of Chicken lysozyme and Baboon alpha-lactalbumin

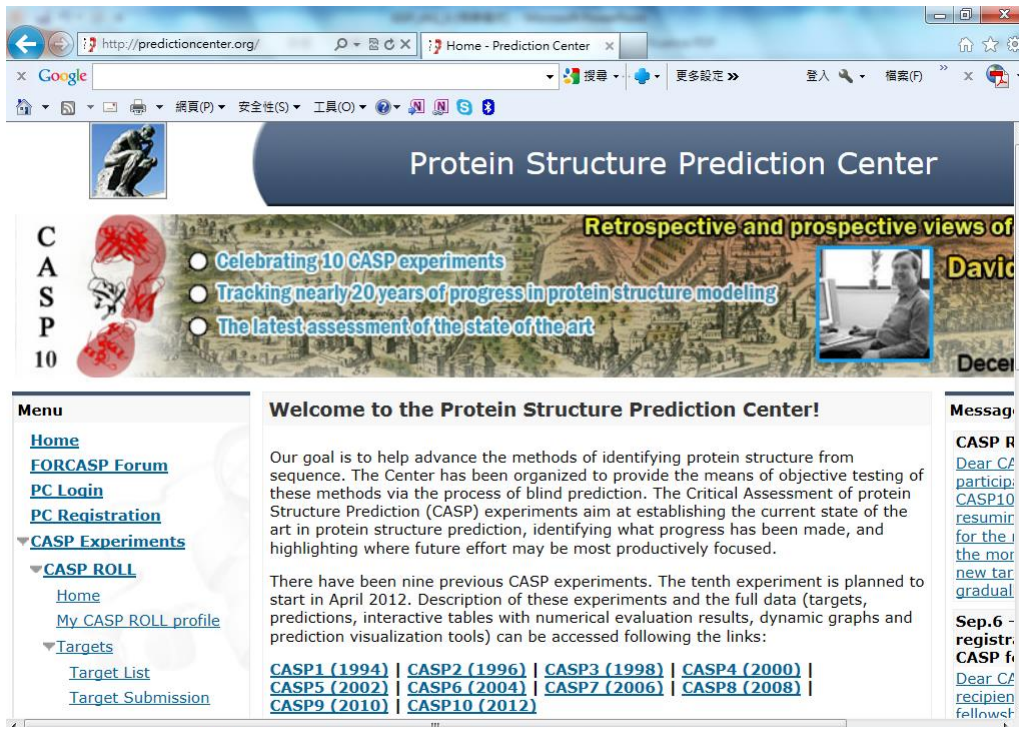
Chicken lysozyme	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGS
Baboon alpha-lactalbumin	KQFTKCELSQNLY-DIDGYGRIALPELICTMFHTSGYDTQAIVEND-ES
Chicken lysozyme	TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS
Baboon alpha-lactalbumin	TEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKILD
Chicken lysozyme	DGN-GMNAWVAWRNRCKGTDVQA-WIRGCRL-
Baboon alpha-lactalbumin	I-KGIDYWIAHKALC-TEKL-EQWL-CE-K

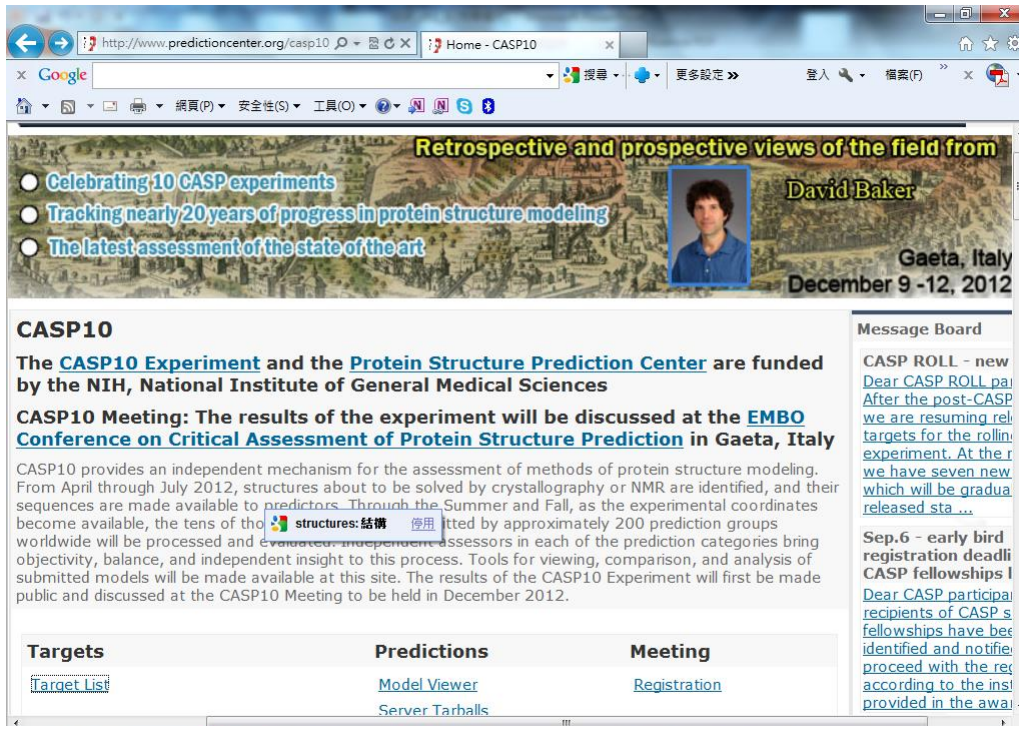
Superposition of Chicken lysozyme (black) and Baboon alpha-lactalbumin (red):

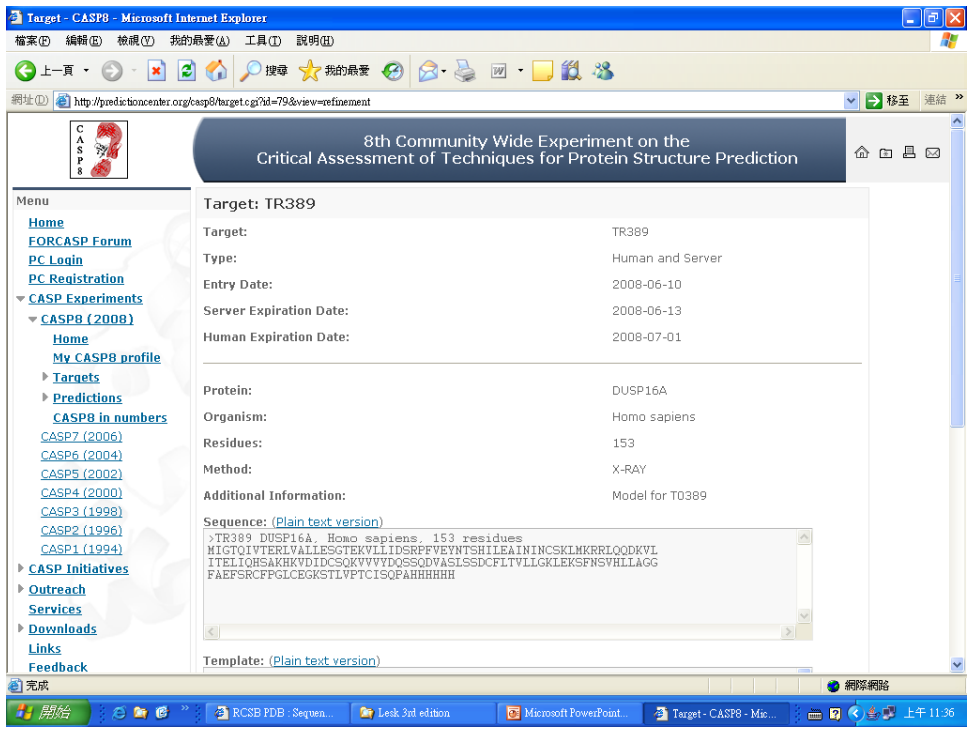


Critical Assessment of Structure Prediction (CASP)

- Judging of techniques for predicting protein structures requires blind test.
- Predictors submit models, which are held until the deadline for release of the experimental structure.
- Then the predictions and experiments are compared – to the delight of a few and the chagrin of most.

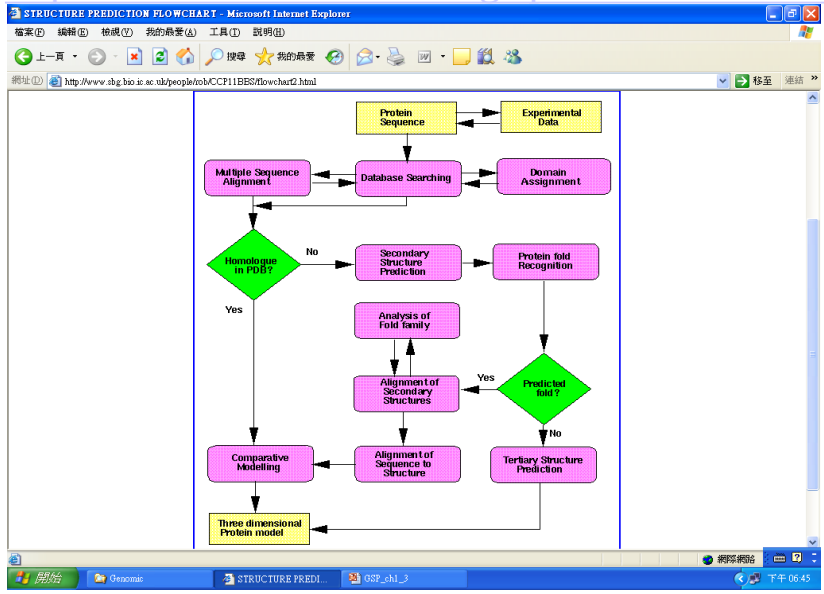



http://www.predictioncenter.org/casp10
Home - CASP10
Retrospective and prospective views of the field from
Celebrating 10 CASP experiments
Tracking nearly 20 years of progress in protein structure modeling
The latest assessment of the state of the art
David Baker
Gaeta, Italy
December 9 -12, 2012
CASP10
The CASP10 Experiment and the Protein Structure Prediction Center are funded by the NIH, National Institute of General Medical Sciences
CASP10 Meeting: The results of the experiment will be discussed at the EMBO Conference on Critical Assessment of Protein Structure Prediction in Gaeta, Italy
CASP10 provides an independent mechanism for the assessment of methods of protein structure modeling. From April through July 2012, structures about to be solved by crystallography or NMR are identified, and their sequences are made available to predictors. Through the Summer and Fall, as the experimental coordinates become available, the tens of thousands of models submitted by approximately 200 prediction groups worldwide will be processed and evaluated. Independent assessors in each of the prediction categories bring objectivity, balance, and independent insight to this process. Tools for viewing, comparison, and analysis of submitted models will be made available at this site. The results of the CASP10 Experiment will first be made public and discussed at the CASP10 Meeting to be held in December 2012.
Message Board
Targets
Target List
Predictions
Model Viewer
Server Tarballs
Meeting
Registration


Target - CASP8 - Microsoft Internet Explorer
http://predictioncenter.org/casp8/target.cgi?id=79&view=refinement
8th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction
Menu
Home
FORCASP Forum
PC Login
PC Registration
CASP Experiments
CASP8 (2008)
Home
My CASP8 profile
Targets
Predictions
CASP8 in numbers
CASP7 (2006)
CASP6 (2004)
CASP5 (2002)
CASP4 (2000)
CASP3 (1998)
CASP2 (1996)
CASP1 (1994)
CASP Initiatives
Outreach
Services
Downloads
Links
Feedback
Target: TR389
Target: TR389
Type: Human and Server
Entry Date: 2008-06-10
Server Expiration Date: 2008-06-13
Human Expiration Date: 2008-07-01
Protein: DUSP16A
Organism: Homo sapiens
Residues: 153
Method: X-RAY
Additional Information: Model for T0389
Sequence: (Plain text version)
>TR389 DUSP16A, Homo sapiens, 153 residues
MIGTQIVTERLVALLESGTEKVLLIDSRPFVEYNTSHILEAININCSKLMKRRLQQDKVL
ITELIQHSAKHKVDIDCSQKVVVVDQSSQDVASLSSDCFLTVLLGKLEKSFNSVHLLAGG
FAEFSRCFPGLCEGKSTLVPTCISQPAHHHHHH
Template: (Plain text version)

STRUCTURE PREDICTION FLOWCHART

http://www.russell.embl.de/gtsp/flowchart2.html



A computer game to learn protein folding

- Maintained by University of Washington, Department of Computer Science
- http://fold.it/
- Learn to play this game and get a score as high as you can
  - Download the “get started”
  - Register an account

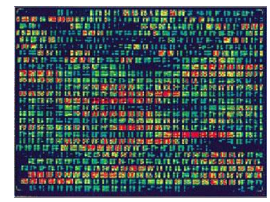
Protein Engineering

- In the laboratory we can manipulate nucleic acids and protein at will.
  - We can probe them by exhaustive mutation to see the effects on function.
  - We can endow (赋予) old proteins with new functions, as in the development of catalytic (催化作用) antibodies
  - We can even create new ones. Engineered proteins must obey the laws of physical chemistry but not the constraints of evolution. With engineered proteins we can explore new territory.

Proteomics

- Combines the census (統計數), distribution, interactions, dynamics, and expression patterns of the proteins within living systems.
- A data-intensive subject, depending on high-throughput measurements
  - Include DNA microarrays, and mass spectrometry.

DNA Microarrays



- Or DNA chips
- Devices for checking a sample simultaneously for the presence of many sequences
- Can be used
  - To determine expression patterns of different proteins by detection of mRNAs
  - For genotyping(遺傳型), by detection of different variant gene sequences, including but not limited to single-nucleotide polymorphisms (SNPs)

Applications of DNA microarrays

- Identifying genetic individuality in tissues or organisms, or genotyping
- Investigating cellular states and processes
- Diagnosis of genetic disease
- Diagnosis of infectious disease
- Specialized diagnosis of disease
- Genetic warning signs
- Drug selection
- Target selection for drug design
- Pathogen (病原體) resistance
- Measuring temporal variations in protein expression

System biology

- Integration – to put all cell part back together
- First aspect:
  - The study of patterns within a cell or an organism: pathways and control cascades, and patterns of protein expression.
  - Patterns have both static and dynamic aspects
    - Identification of pairs of proteins that bind to each other, and assembly of pairwise interactions into a network Static pattern.
    - Dynamic pattern: the flow of metabolites through a network of enzymes, or the flow of information down a control cascade, is a dynamic pattern.

- Second aspect – comparison of occurrence, activities and interactions of genes and proteins ***across different species***.
  - The systems we are trying to understand arose through processes of evolution. Different species illuminate one another.
- High-throughput methods of genomics and proteomics provide data about sequences, expression patterns and interactions.
  - Systems biology takes the data as pieces of a jigsaw puzzle that extends in both space and time. To understand the complex and delicate instrument that is the living cell, we must fit the pieces into their frame.

Clinical implications

1. Diagnosis of disease and disease risks
   - DNA sequencing can detect the absence of a particular gene, or a mutation.
2. Genetics of responses to therapy – customized treatment
   - People differ in their ability to metabolize drugs, different patients with the same condition may require different dosages.

3. Identification of drug targets

A target is a protein the function of which can be selectively modified by interaction by a drug, to affect the symptoms or underlying causes of a disease.

4. Gene therapy

If a gene is missing or defective, we'd like to replace it or at least supply its product. If a gene is overactive, we'd like to turn it off.

Direct supply of proteins is possible for many diseases.

Practice

- Huntington disease
- Find out the cause of this disease using the Internet search.
- What is the phenomenon of "anticipation"?
- Answer:
- The same questions for other diseases: 地中海型貧血 Mediterranean anemia，紅斑性狼瘡Systemic Lupus Erythematosus