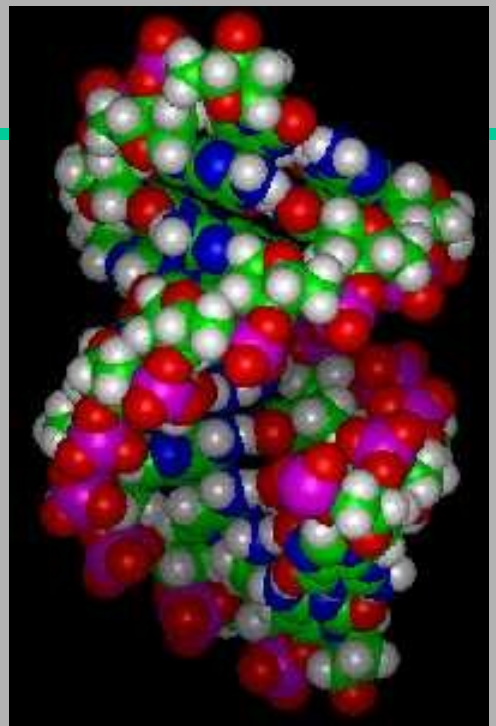
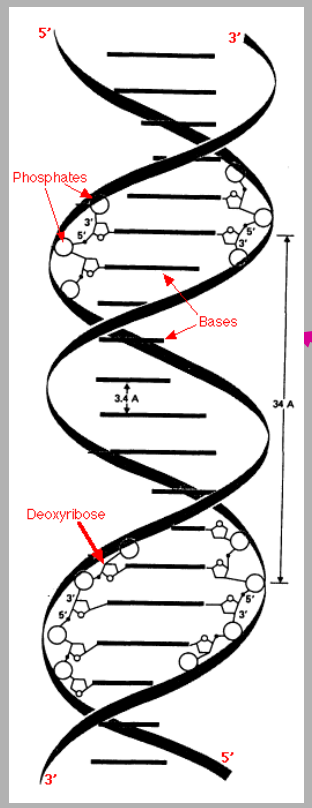
Bioinformática
Visión Global…

Tópicos Selectos en Bioinformática y Biotecnología

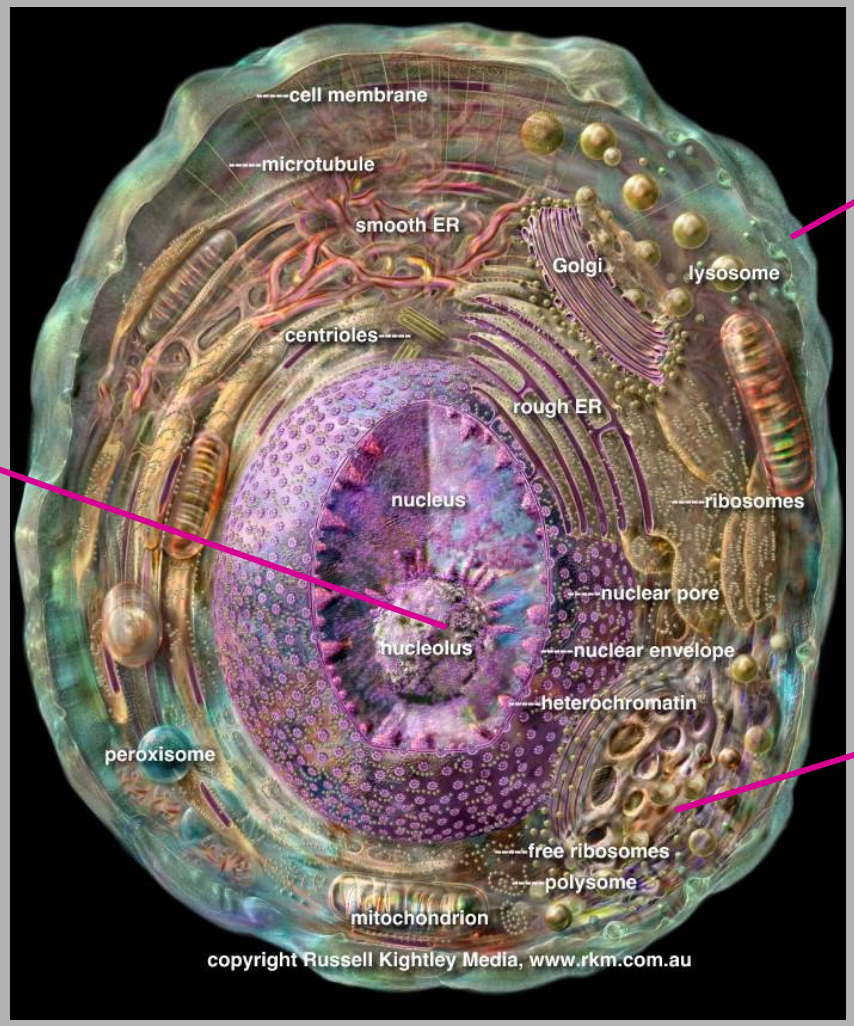


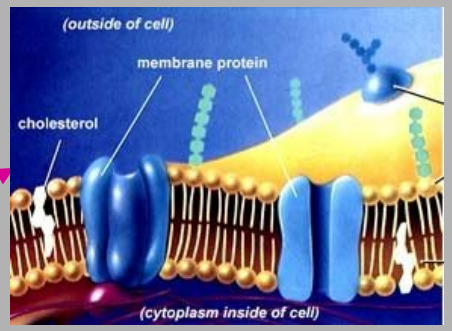
Mirko Zimic

La Biología está basada en la materia ´suave´ viviente

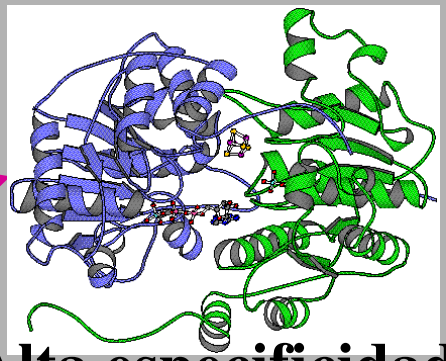


Información



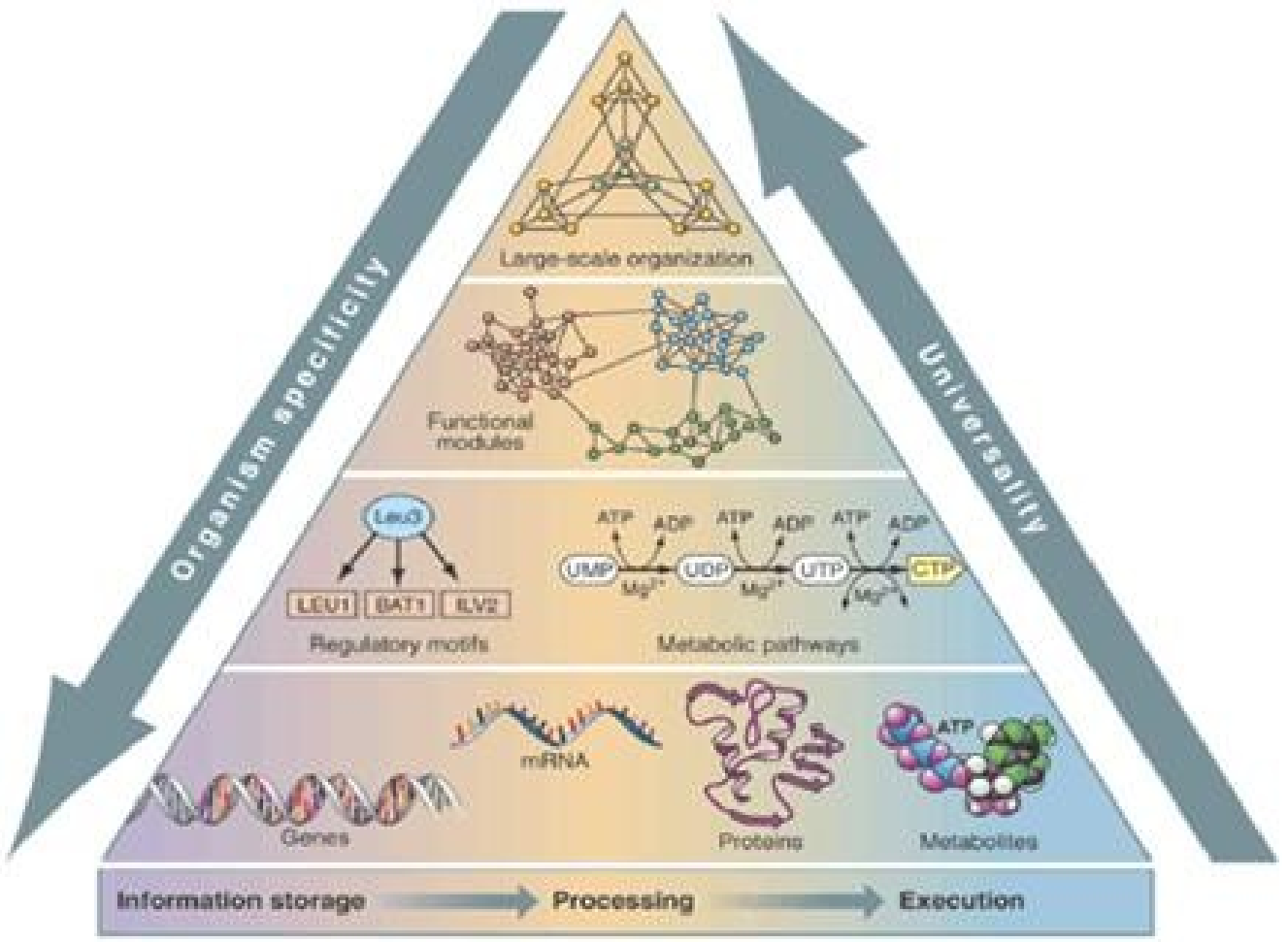


Auto-ensamblaje



Alta especificidad

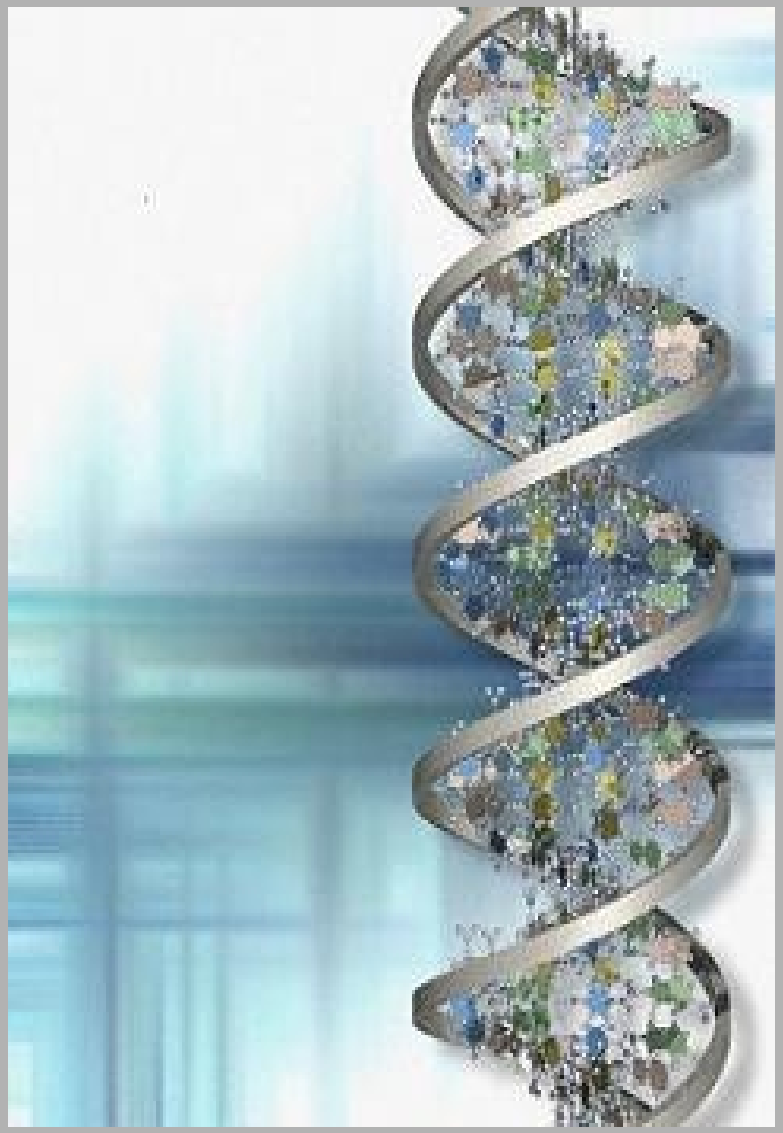
*"**Los objetos vivientes** están compuestos por moléculas inertes"*
Albert Lehninger



El problema es:

Cómo estas moléculas confieren la admirable combinación de características que denominamos vida???

Cómo es que un organismo vivo aparece ser más que la suma de sus partes inanimadas???



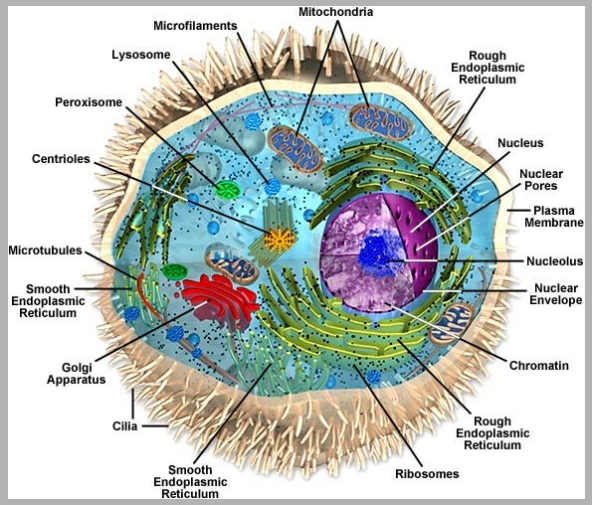
La Física procura entender y reducir la Biología en leyes fundamentales

Pero este es un problema muy complicado !

Son demasiadas las variables y resulta imposible describir un sistema de un número tan grande de partículas

SOLUCIÓN:
scripción estadística del mundo ´aleator

Si todo en el nano-mundo de las células es aleatorio,
cómo podemos realizar predicciones?



n ACTIVIDAD COLECTIVA de ‘muchos’ objetos de
ovimiento aleatorio puede ser predicho, aun cuando e
ovimiento exacto de un sólo objeto es desconocido

What is Bioinformatics?

- **What is Bioinformatics? -** Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- **What is Computational Biology? -** The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

(Working Definition of Bioinformatics and Computational Biology - July 17, 2000).
http://www.grants2.nih.gov/grants/bistic/CompuBioDef.pdf

The “Ideal” Syllabus

- **Molecular Biology**

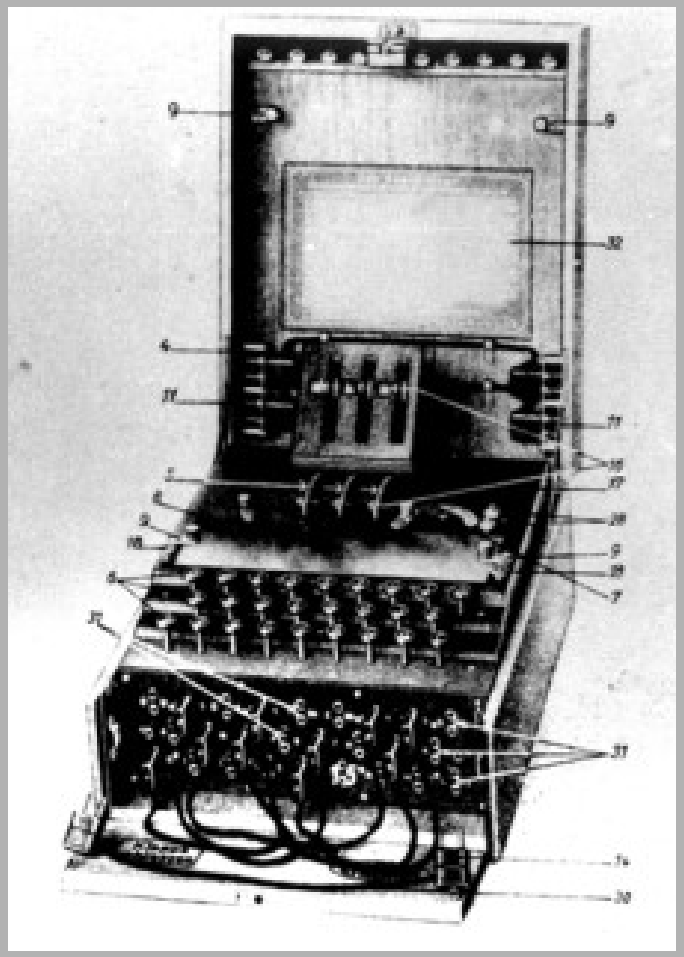
  Basic concepts, Genomic and Proteomic structure
- **Core Bioinformatics**

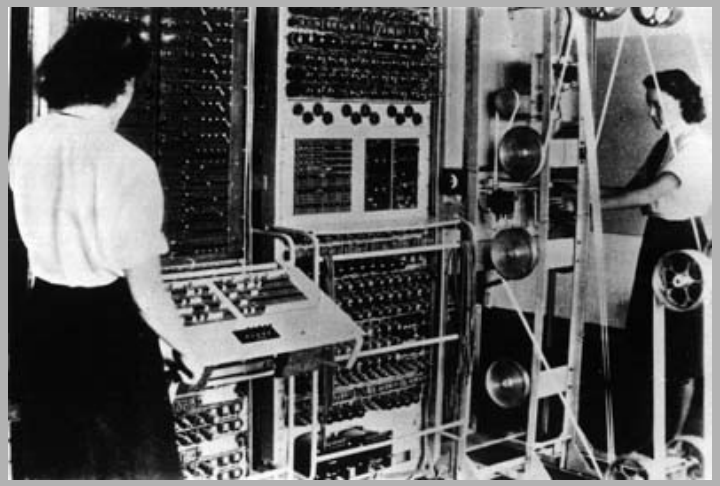
  Biological Databases, Sequence Analysis,
  Functional Genomics
- **Advanced Bioinformatics**

  Molecular Evolution and Phylogeny
  Protein Structure Prediction
  The Transcriptome
  The Proteome
- **Informatics**

  Information Theory
  Basic Statistics
  Database Technologies
  Knowledge Representation
  Biocomputing

Durante la II Guerra mundial los ingleses construyen en respuesta al codificador Enigma, el Colossus.

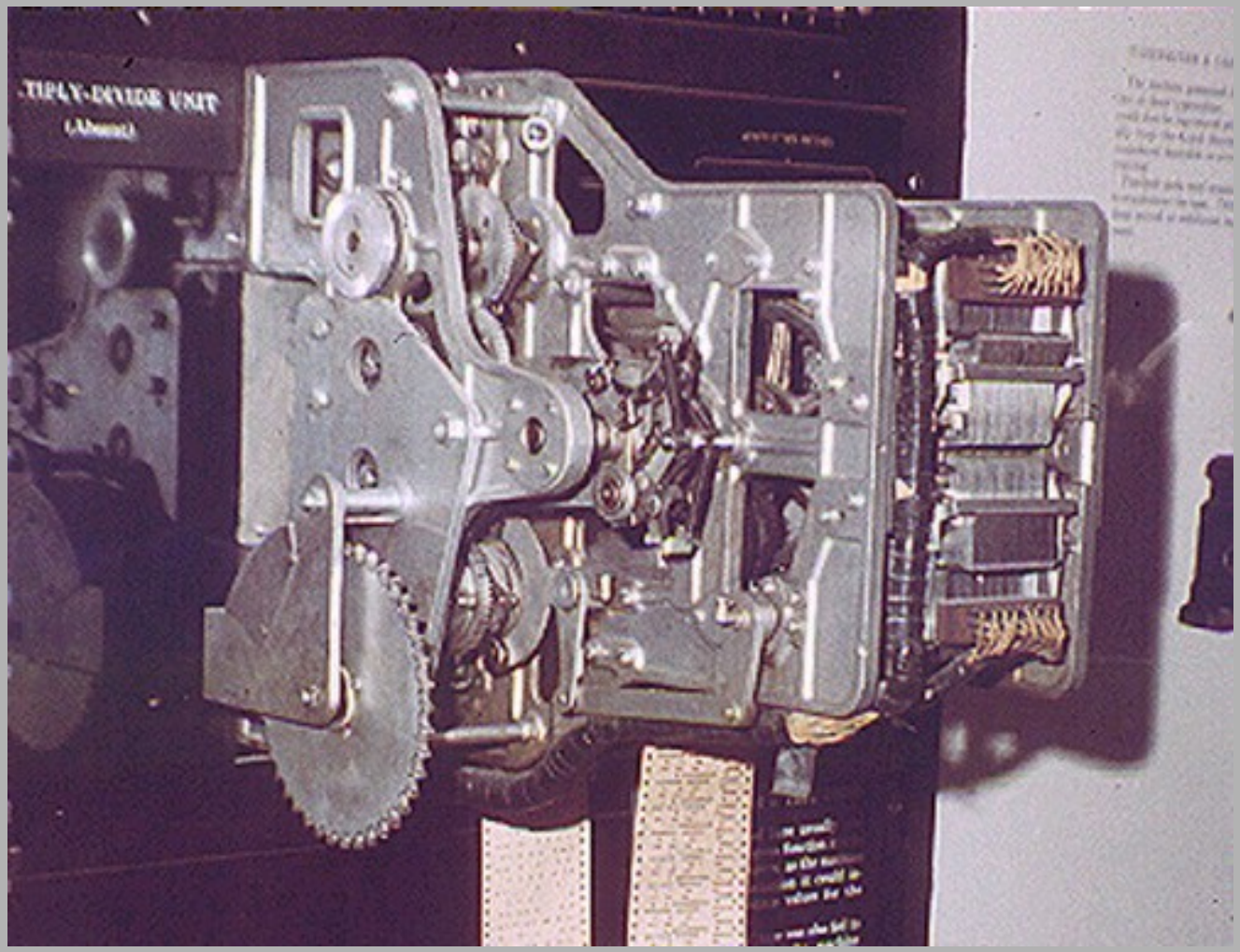




Enigma

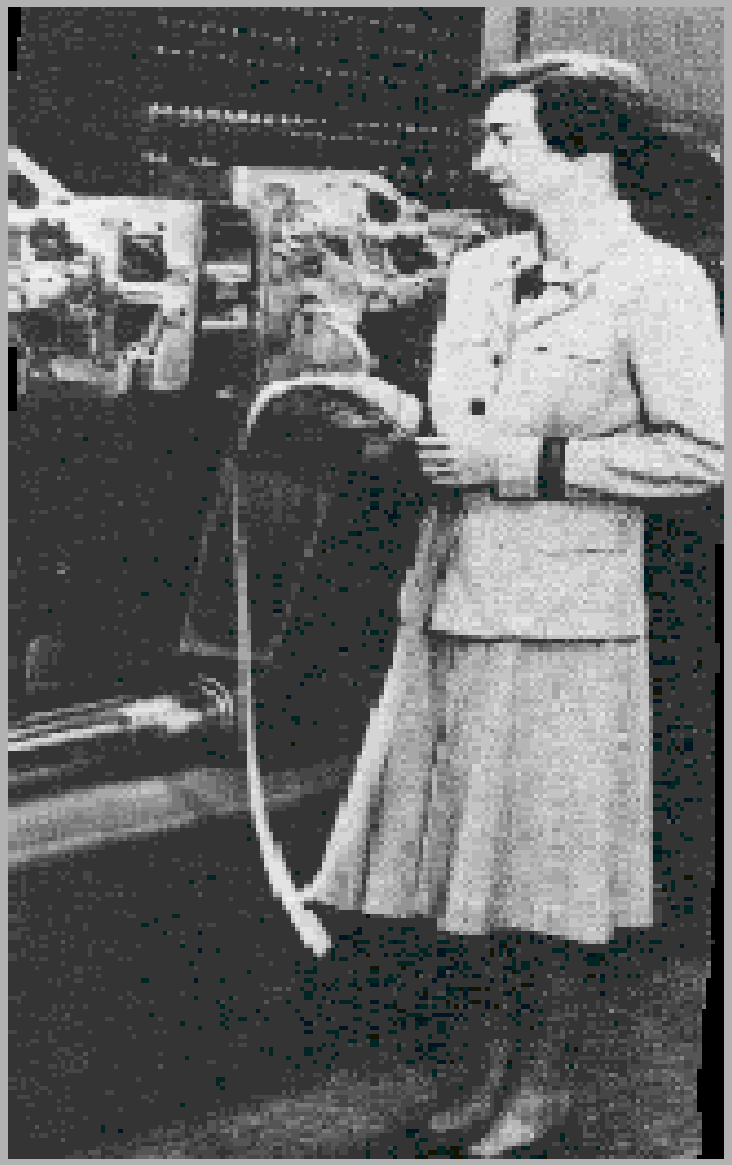
Konrad Zuse con la Z1 reconstruída. Zurich

En 1944 IBM y la Universidad de Harvard estrenan Mark I, la primera computadora que responde a la moderna definición. Medía.15 metros de largo, 2.40 mts de alto y pesaba 10 toneladas. Utilizaba relays electromecánicos.



Este es uno de los relay que se usaron en la Mark I

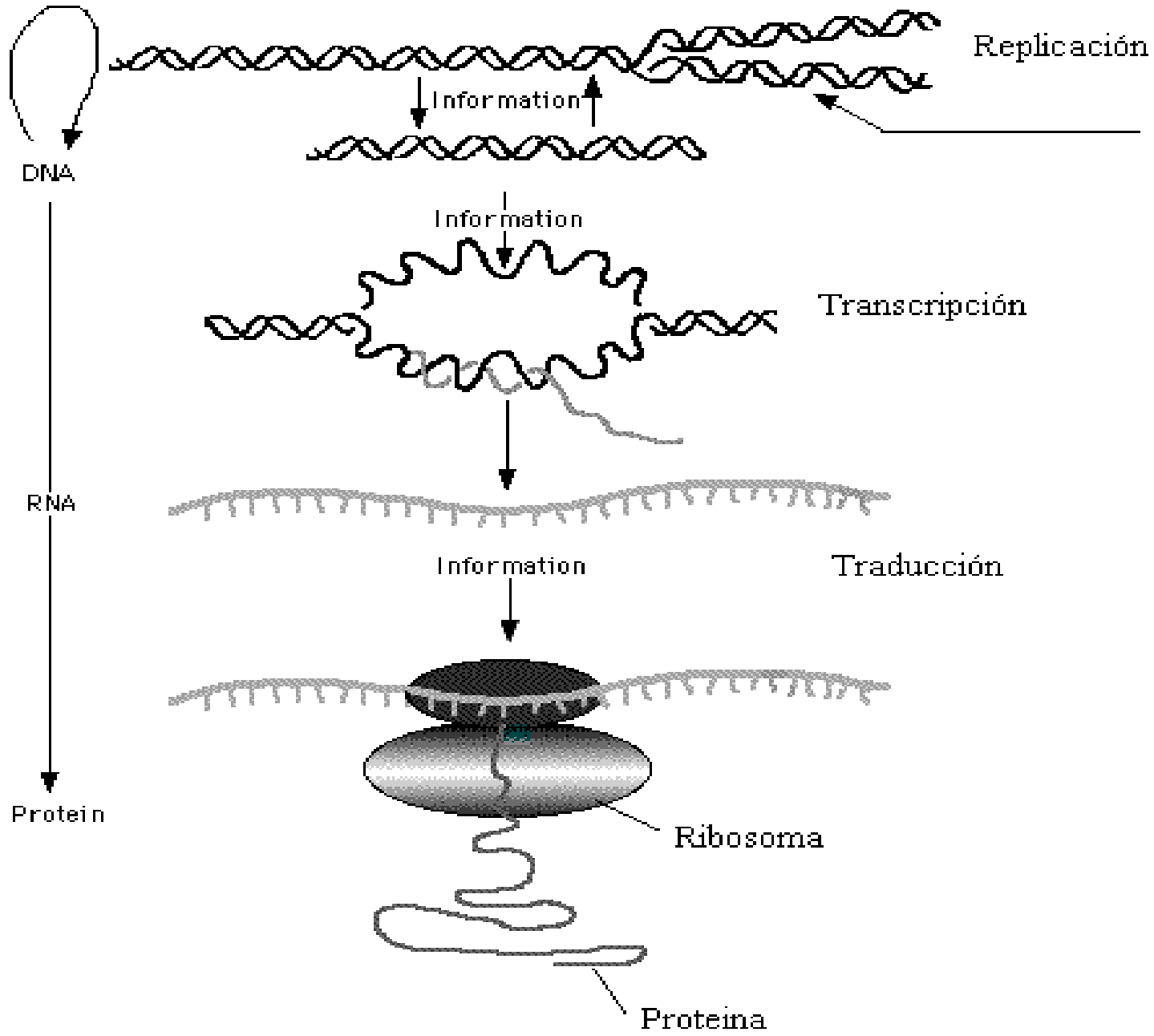
Sumaba en menos de un segundo, multiplicaba en cerca de seis, y dividía en cerca de doce.



Costo Efectividad !

La Bioinformática resulta ser una disciplina muy favorable en cuanto a costo-efectividad.

El dogma central de la Biología Molecular



Collecting Sequence Data

- **Genome (DNA-level): Genomic sequencing**
  - Complete picture of genome
  - Generates physical map
  - Includes regulatory and other silent regions
- **Transcriptome (RNA-level): Expression-library sequencing**
  - Expressed genes only
  - Splicing / variant forms
  - Can correlate with levels of expression
- **Proteome (protein-level): Protein sequencing**
  - Insight into biological function
  - Gives information on protein-protein interactions
  - Post-translational modifications detected

- La Bioinformática se inicia con el desarrollo de bases de datos biológicas, seguido del desarrollo de herramientas de búsqueda rápida de información…

- Actualmente la Bioinformática busca el desarrollo de algoritmos de predicción basado en la información almacenada en las bases de datos biológicas.

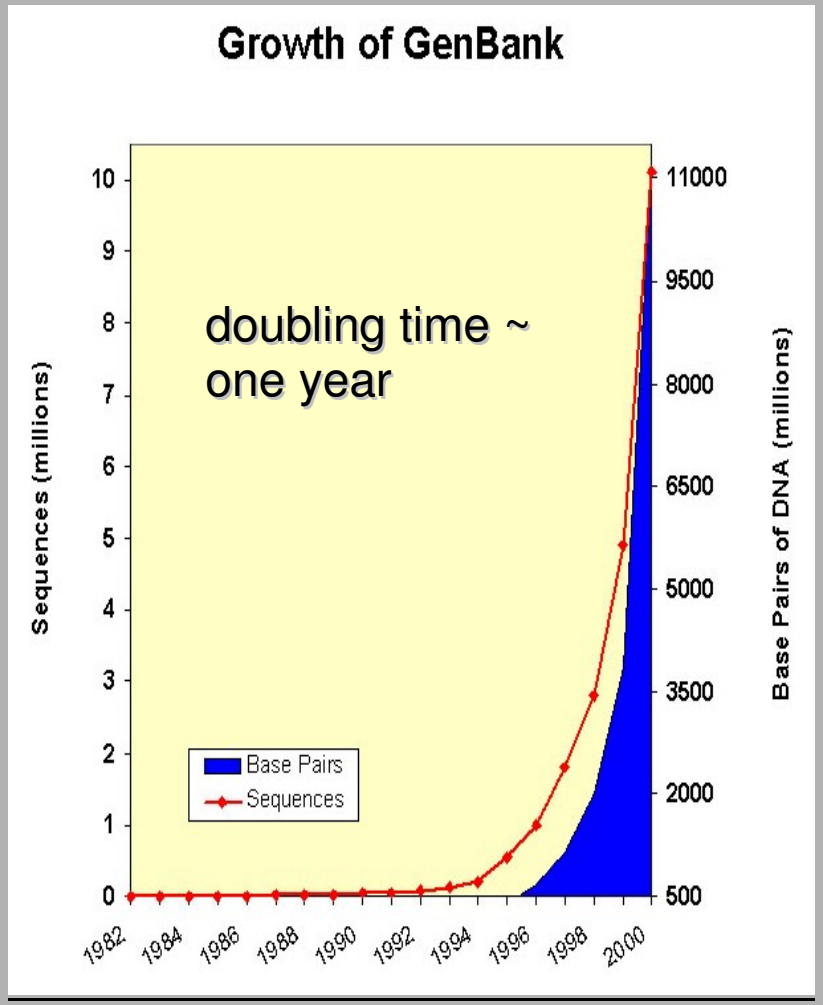
Historical Perspective

Key developments:

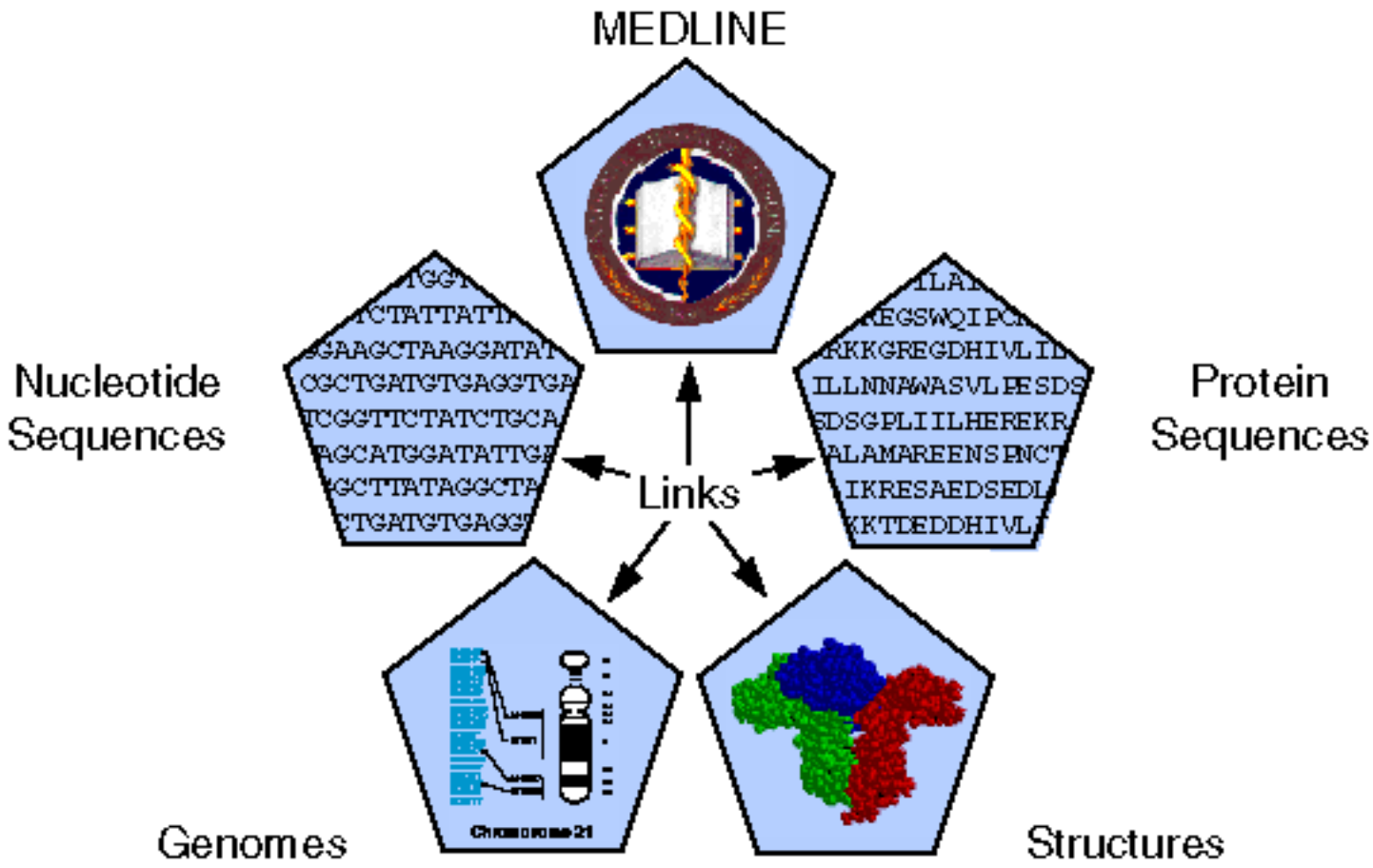
- Dayhoff, *Atlas of Protein Sequence and Structure* (1965-1978)
- Genbank/EMBL nucleic-acid sequence databases (1979-1992)
- Entrez (early 90's – date)
- Sequence alignment algorithms: Needleman/Wunsch (1970), Smith/Waterman (1981), FASTA (Pearson/Lipman, 1988), BLAST (Altschul, 1990)
- Genomes (1995 – date)

The exponential growth of molecular sequence databases & cpu power —

Year	BasePairs	Sequences
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	11101066288	10106023
2001	14396883064	13602262



Databases contain more than just DNA & protein sequences



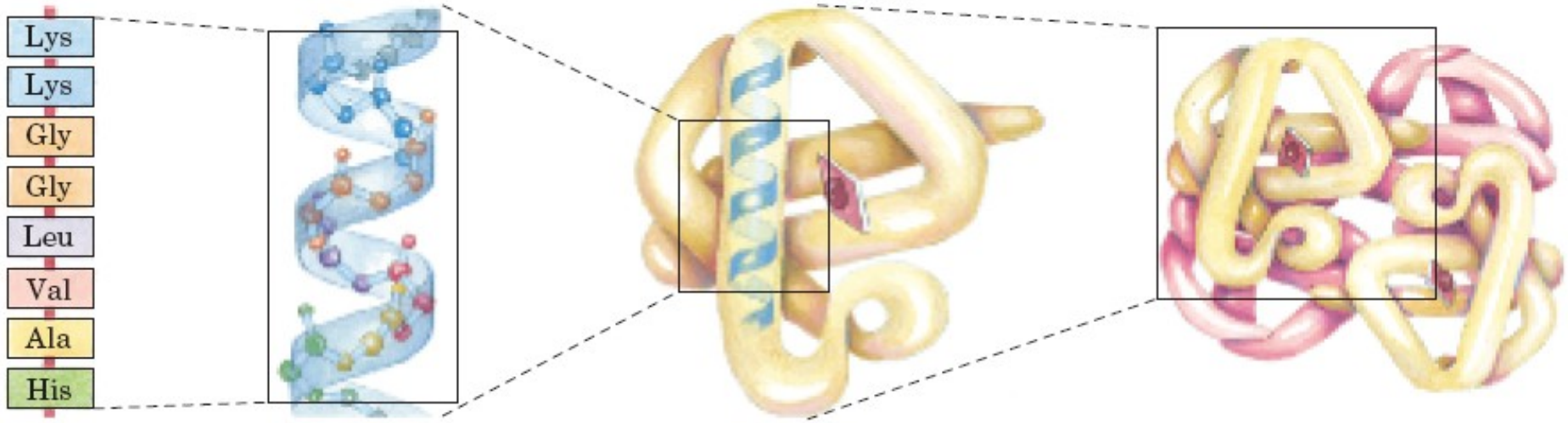
The "*omics*" Series

- **Genomics**
  - **Gene identification & charaterisation**
- **Transcriptomics**
  - **Expression profiles of mRNA**
- **Proteomics**
  - **functions & interactions of proteins**
- **Structural Genomics**
  - **Large scale structure determination**
- **Cellinomics**
  - **Metabolic Pathways**
  - **Cell-cell interactions**
- **Pharmacogenomics**
  - **Genome-based drug design**

Structural Genomics

What is structural genomics?

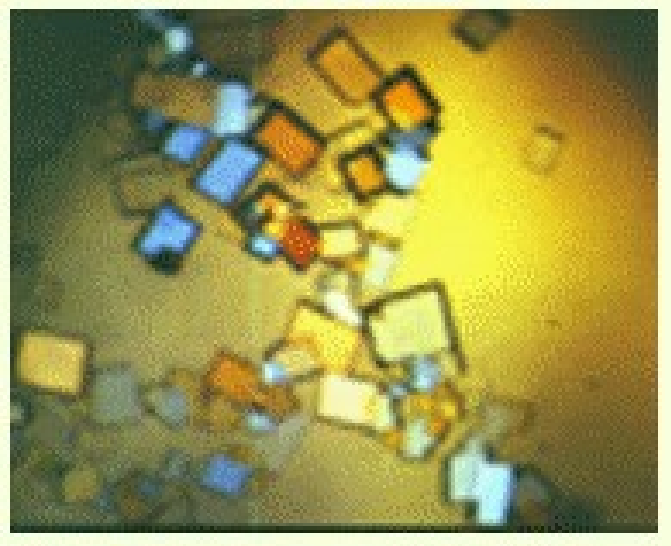
- Genomes and folds:
  - Finding folds in genomes
  - Structural properties of entire proteomes
  - Comparing genomes in terms of structure
- Selection of targets for structural genomes
  - Covering the sequence space with structures
  - Using structure to understand function
  - Systematic structure determination for complete genomes
  - Special targets
  - Predicting success of structure determination
- Adaptation of proteins to extreme environments
- Structural genomics resources on the internet

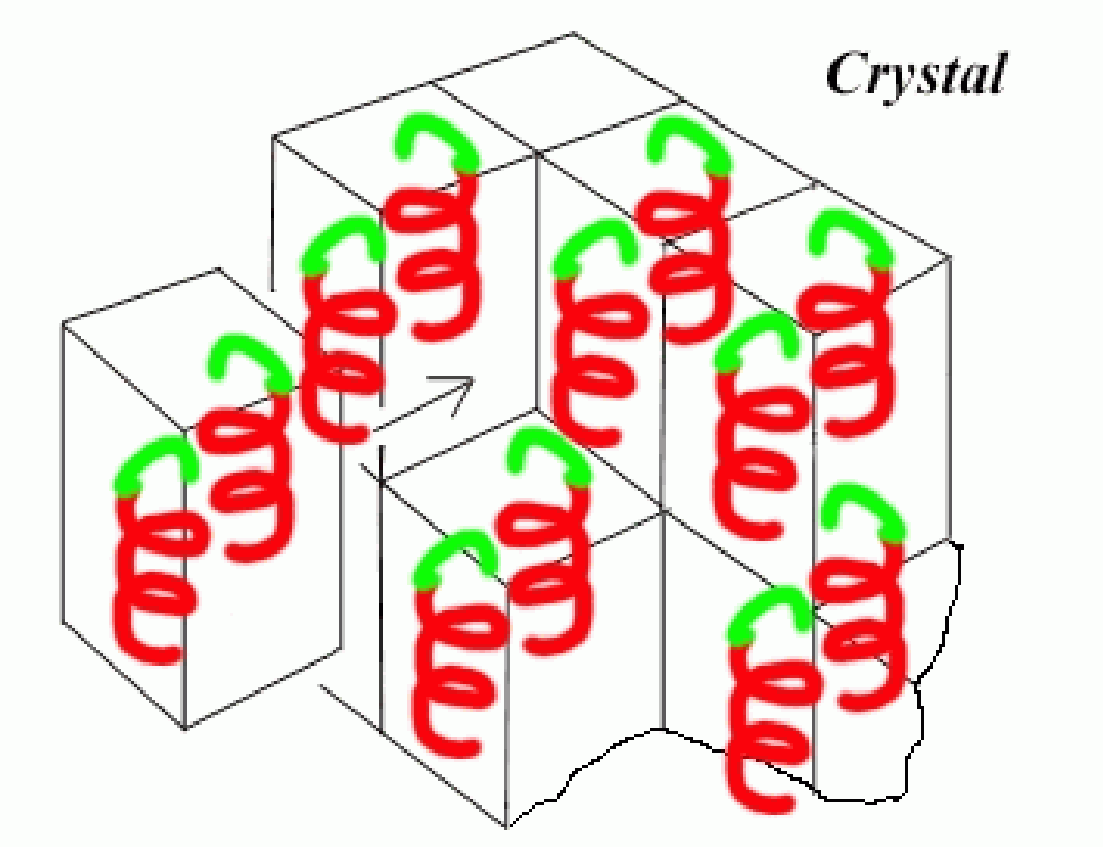


Estructura 3D de las proteínas

Difracción de rayos X

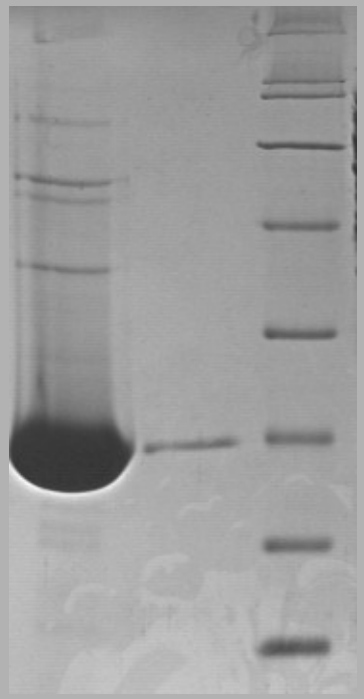
- Determina espaciado interatómico de una red cristalina ordenada (proteína).

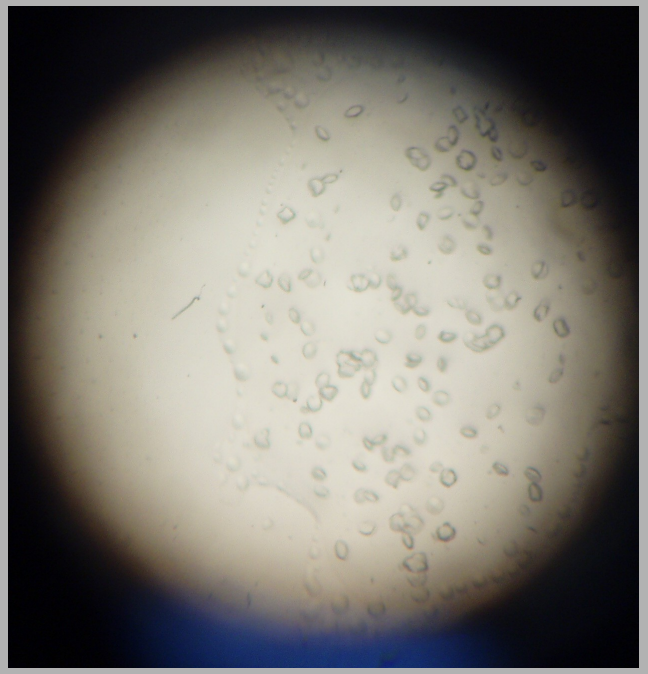


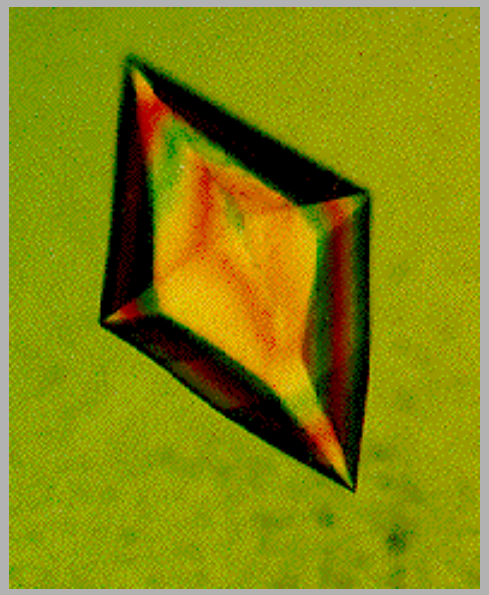


Difracción de rayos X

- Determina posición de átomos dentro de un cristal.



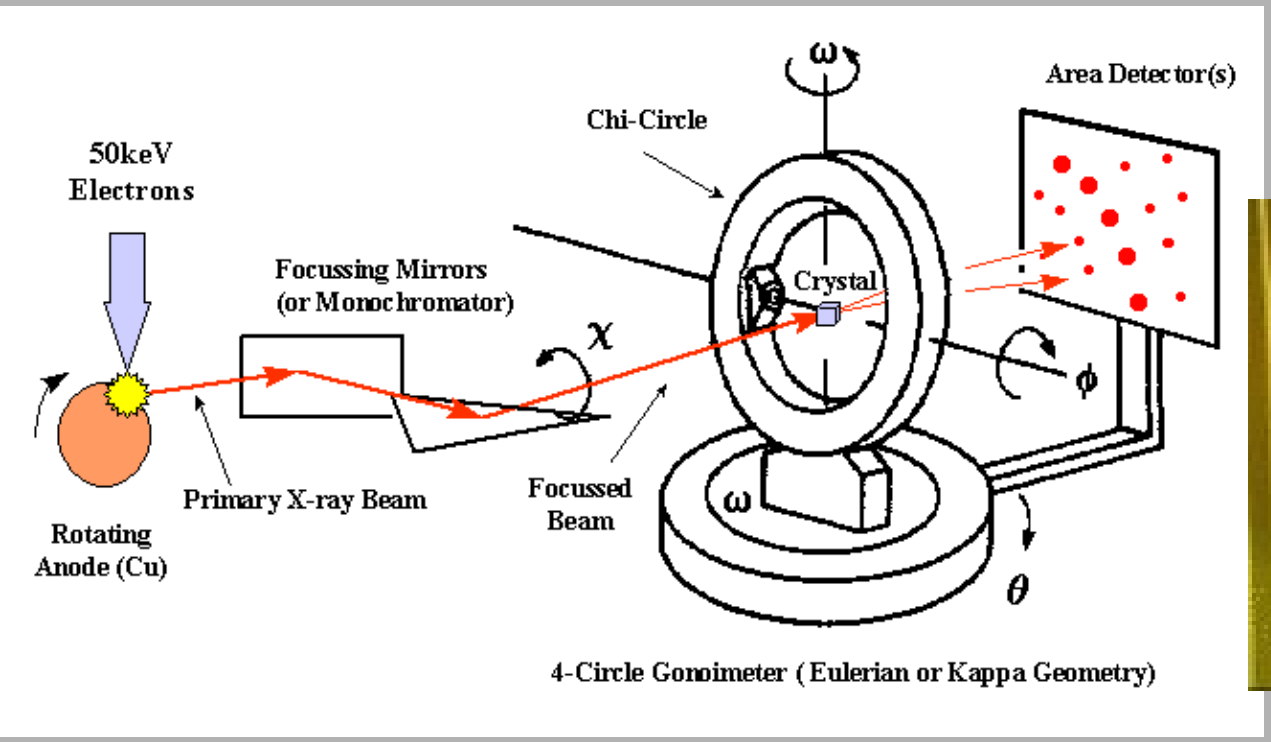




Identificación → Purificación → Cristalización

Difracción de rayos X

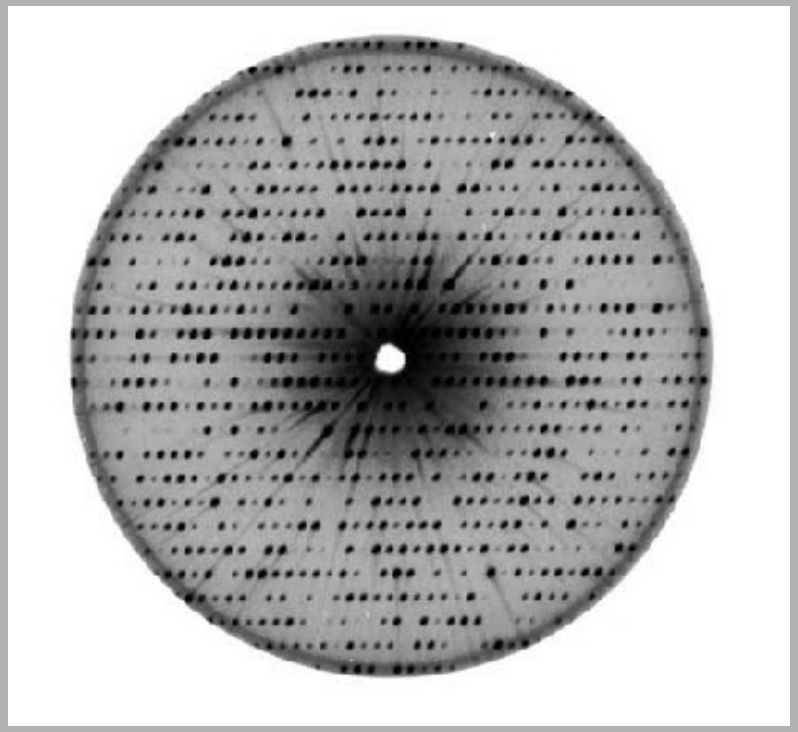
- Rayos X son difractados por e^- de los átomos.



Goniómetro controla la posición del cristal

Difracción de rayos X

- Señales regulares de difracción; transformada de Fourier.

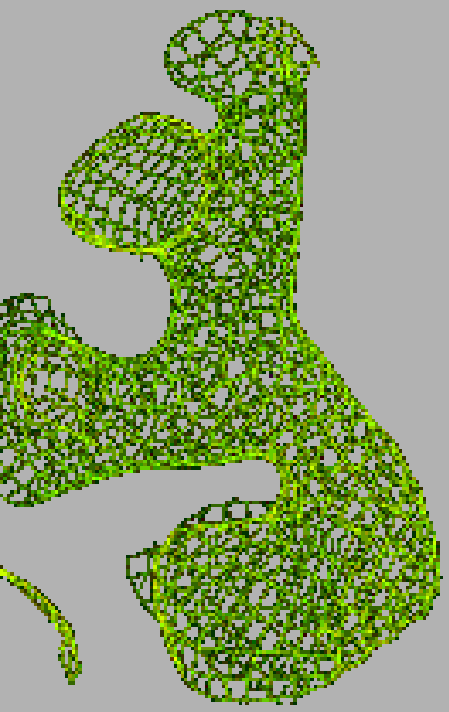


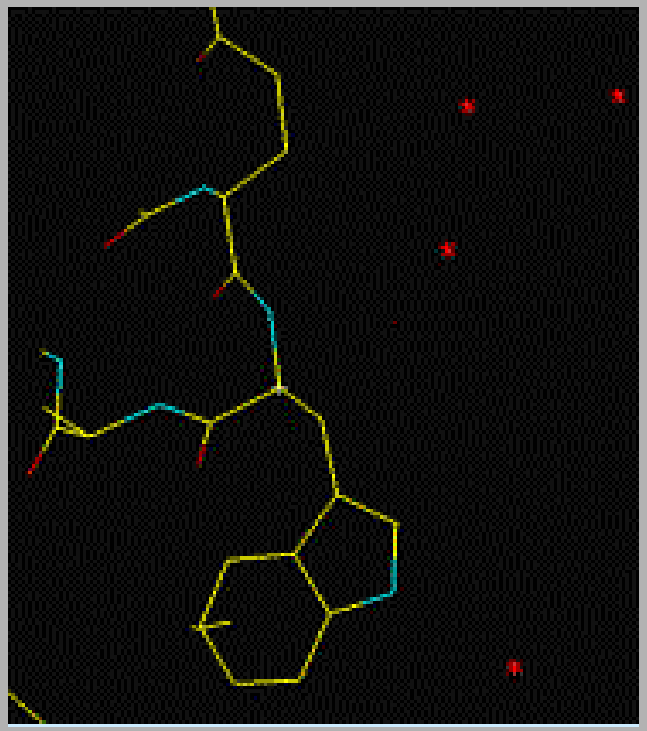
Patrón de difracción

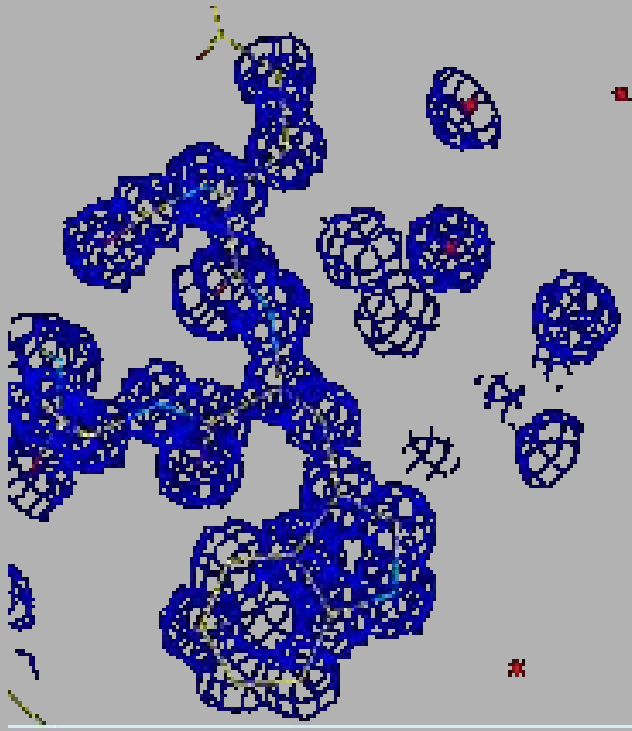
$$F(h) = \int_{-\infty}^{+\infty} f(x) \cdot e^{2\pi i (hx)} dx$$

Difracción de rayos X

- Dan escasa información del movimiento molecular





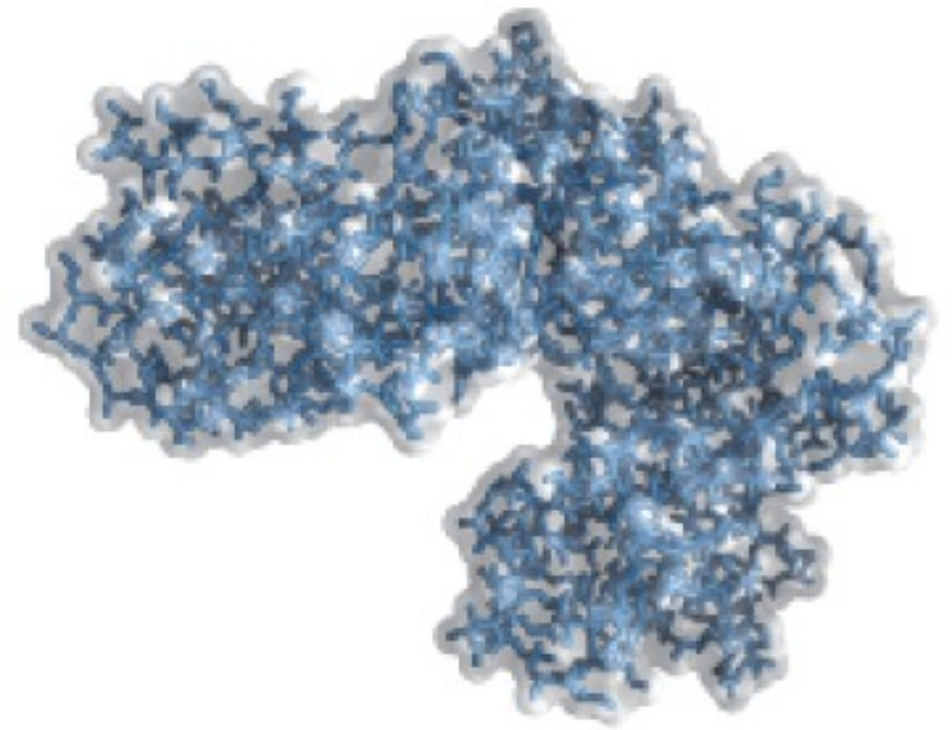


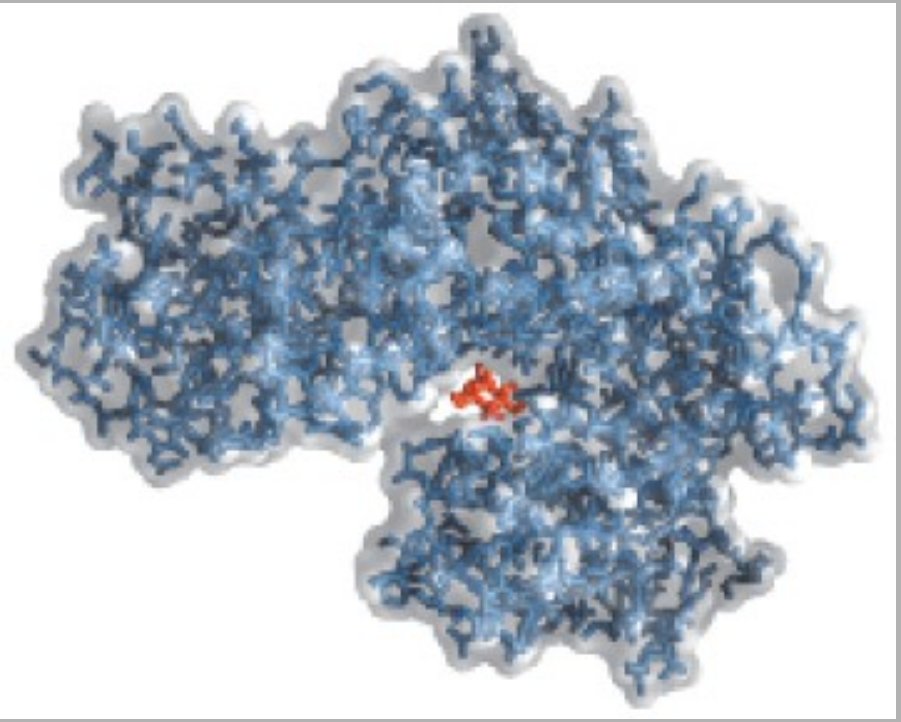
Mapa de densidad electrónica

Resolución y precisión del modelo

Difracción de rayos X

- Casi siempre: conformación funcional de la proteína

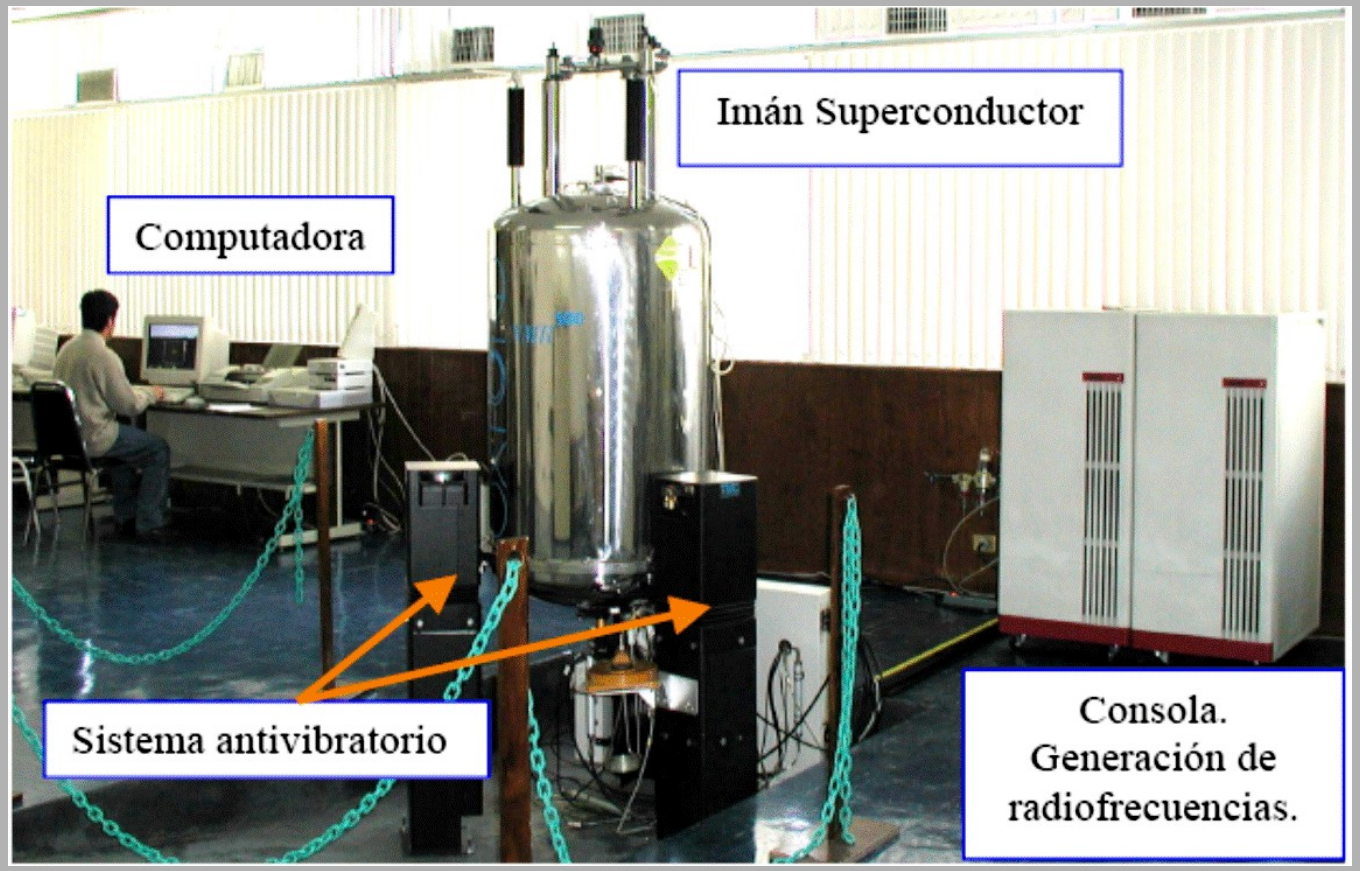




Predicción de cambio conformacional

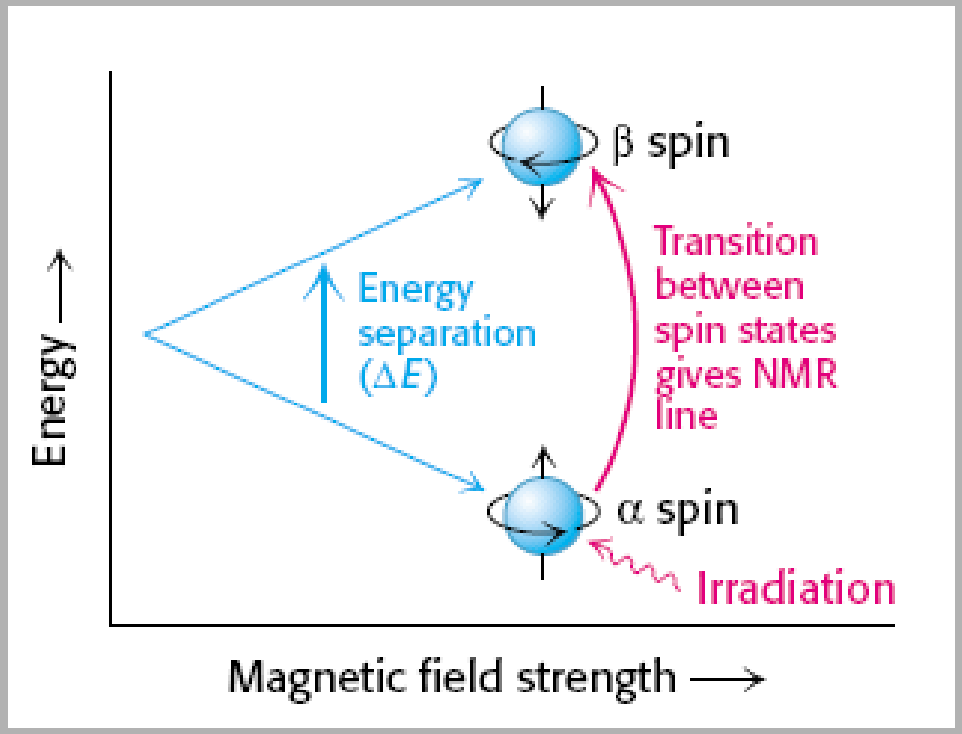
Resonancia magnética nuclear

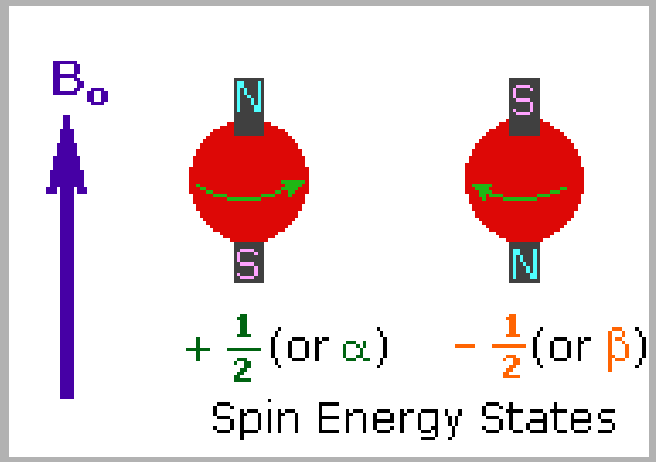
- Revela estructura atómica de moléculas en solución.
- No necesita cristales, sólo proteína pura en solución.



Resonancia magnética nuclear

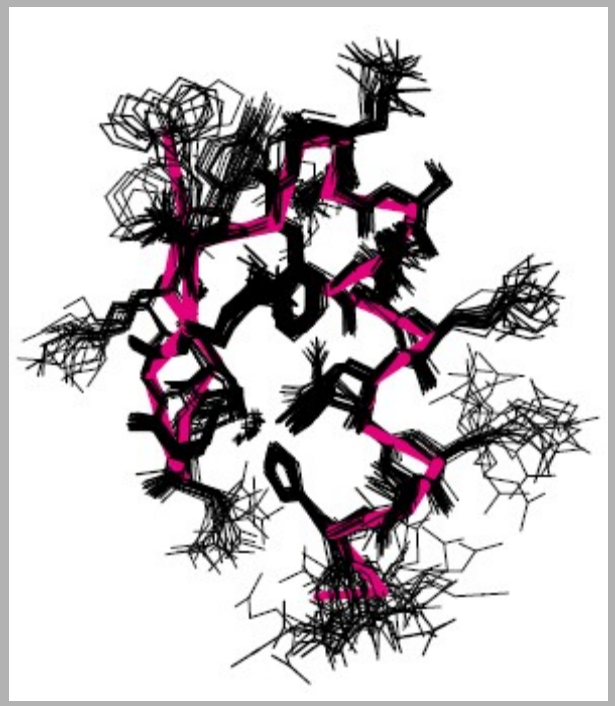
- Depende del hecho que ciertos núcleos atómicos son intrínsecamente magnéticos.

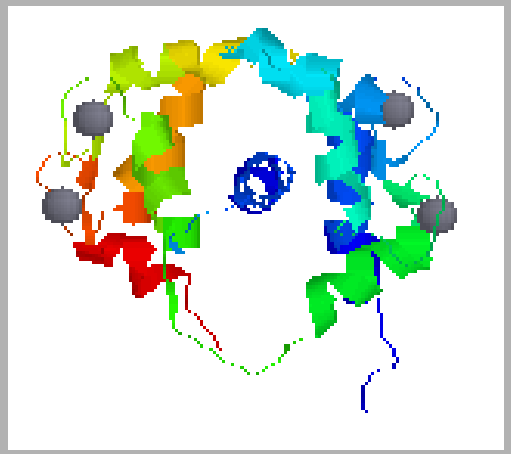


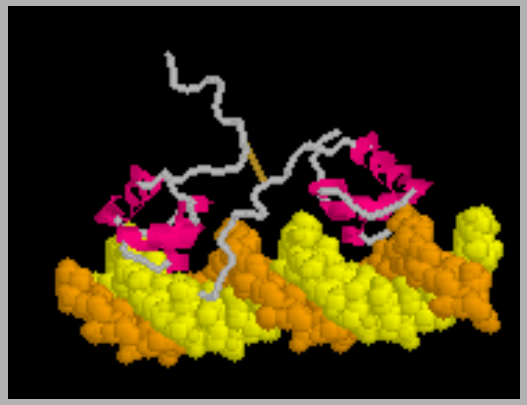


Resonancia magnética nuclear

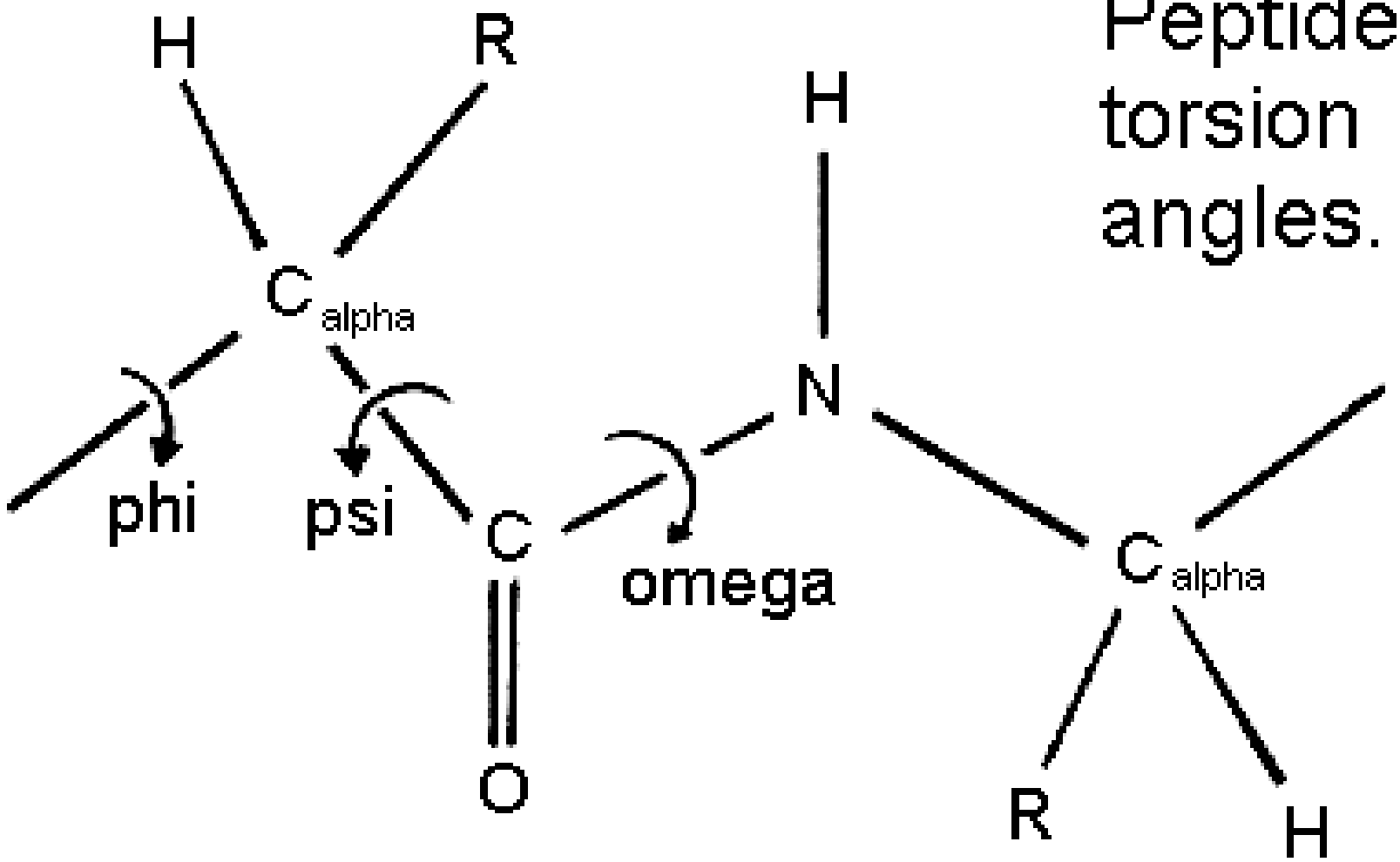
- Se puede estudiar cambios conformacionales, plegamiento, interacciones con otras moléculas.







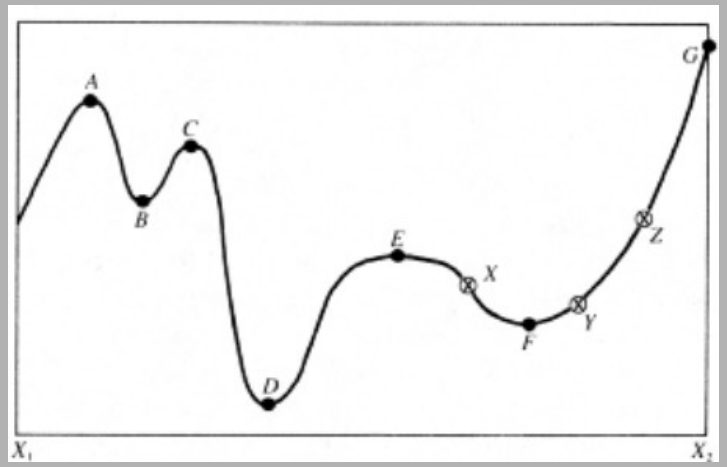
Peptide torsion angles.



Energy minimisation

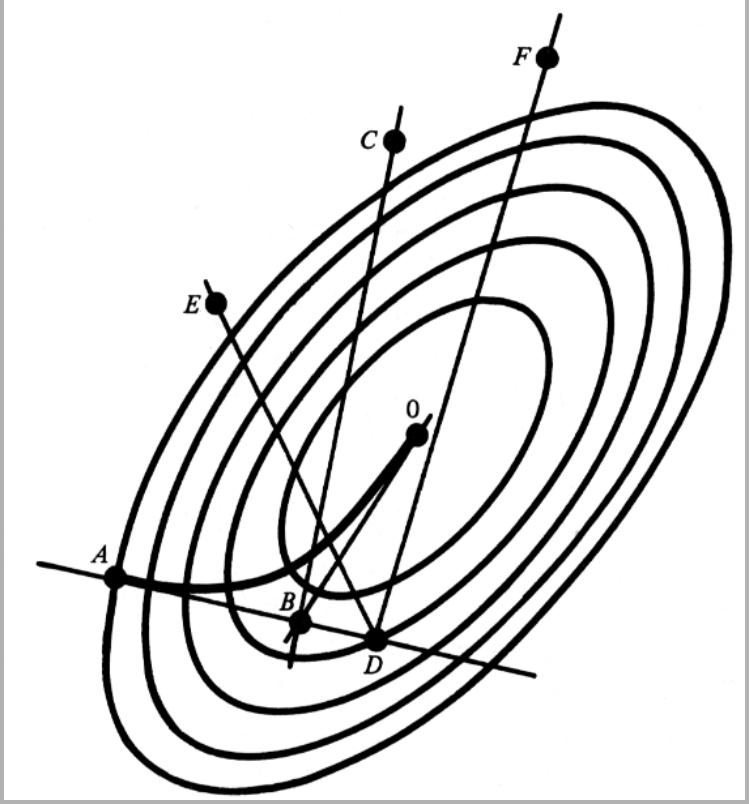
- Calculation of how atoms should move to minimise TOTAL potential energy
- At minimum, forces on every

  atom are zero.
- Optimising structure to remove strain & steric clashes
- However, in general finds local rather than global minimum. Energy barriers are not overcome even if much lower energy state is possible ie structures may be locked in. Hence not useful as a search strategy.



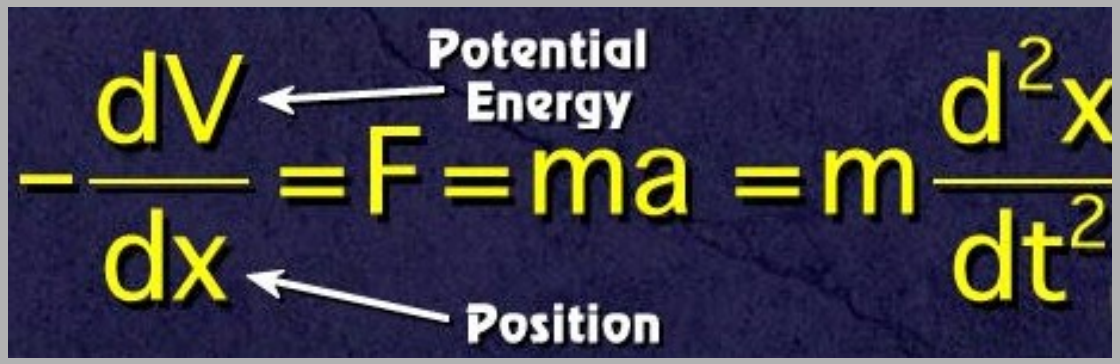
Energy minimisation

- Potential energy depends on many parameters
- Problem of finding minimum value of a function with >1 parameters. Know value of function at several points.
- Grid search is computationally not feasible
- Methods
  - Steepest descents
  - Conjugate gradients



Molecular Dynamics: Introduction

Newton's second law of motion



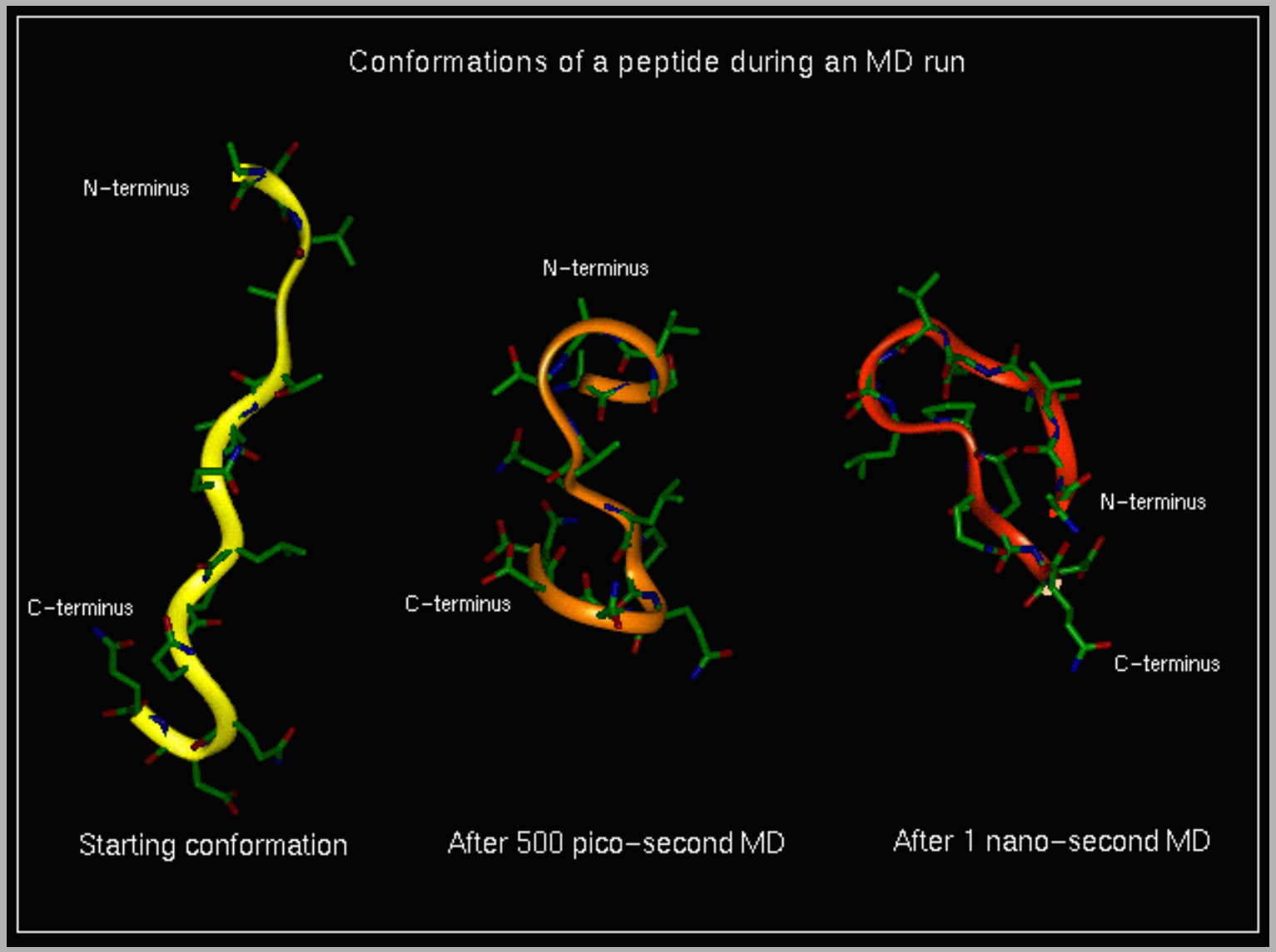
Molecular Dynamics: Introduction

In general, given the values x_1, v_1 and the potential energy $V(x)$, the molecular trajectory $x(t)$ can be calculated, using,

$$x_i = x_{i-1} + v_{i-1}\Delta t$$

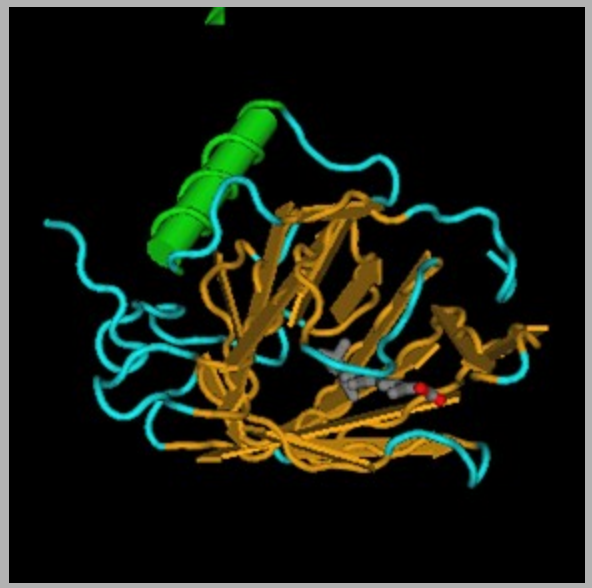
$$v_i = v_{i-1} - m^{-1} \left.\frac{dV(x)}{dx}\right|_{x_{i-1}} \Delta t$$

How a molecule changes during MD



Functional Genomics

- Development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information provided by structural genomics.



retinol-binding protein
(NP_006735)

β -lactoglobulin
(P02754)

Pairwise alignment of retinol-binding protein and β -lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD.   ADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENG CAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV.  ........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFC ENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIV                RQYRLIV 185 RBP
    . |       |     | :    ||   .
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSF                ....... 178 lactoglobulin
```

Identity (bar)

Pairwise alignment of retinol-binding protein and β -lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDET  MSATAKGRVR.LLNNWD..VCADMVGTFT  E 97 RBP
    : | |   |   |     :  | .| .  ||  |:   ||         .
 45 ISLLDAQSAPLRV.YVL LKPTPEGDLEILLQKWENGECAQKKIIAE  K 93 lactoglobulin

 98 DPAKFKMKYWGVASFL GNDDHWIVDTDYDTYAV..........Q  C 136 RBP
     || ||.        |        :.||||  | .             |
 94 IPAVFKIDALNENKVL .....VLDTDYKKYLLFCMENSAEPEQS  C 135 lactoglobulin

137 RLLNLDGTCA  PPEAQKIVRQRQ.EELC  RBP
    . |         |   .      | || |
136 QCLVRTPEVD  PMHIRLSFNPTQLEEQC  lactoglobulin
```

Somewhat similar (one dot)

Very similar (two dots)

Pairwise alignment of retinol-binding protein and β -lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL.......VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFS[illegible]NGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDK[illegible]KALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```

Internal gap

Terminal gap

PRETTYBOX of: prettybox.msf{*} October 26, 1998 11:05:11.41

```
fa10 ..........  ..........  ..........  ..tttgeSa  d.PvtTtVE.  ...N..yggD  21
fa12 ..........  ..........  ..........  ..ttatgeSa  d.PvtTtVE.  ...N..ygge  21
folk ..........  ..........  ..........  ..ttsageSa  d.PvtTtVE.  ...N..ygge  21
   e GvenAe.kgv  tENTna.Tad  fvaqpvyLPe  .nqt......  kv.Affynrs  ...sPi.gaf  46
 plm GLGqMLEsmI  .dNTvrETVG  AATSRDaLPN  TeaSGPtHSK  EIPALTAVET  GATNPLVPSD  59
 pls GLGqMLEsmI  .dNTvrETVG  AATSRDaLPN  TeaSGPaHSK  EIPALTAVET  GATNPLVPSD  59
 p2s GiGDMiEgav  .Egitknalv  ppTStnsLPq  hkpSGPaHSK  EIPALTAVET  GATNPLVPSD  59
 p3s GieDliseva  .qgal..Tls  lpkqqDsLPd  TkaSGPaHSK  EvPALTAVET  GATNPLaPSD  57
 cb3 ...qpvEdaI  .......T..  AAigR..vad  TvgtGPtnSe  aIPALTAaET  GhTsqvVPqD  46
 r14 GLGDeLEevI  vEkT.kqTV.  Asi.......  ..sSGPkHtq  kvPiLTAnET  GATmPvlPSD  49
  r2 ...npvEnyI  dEvlnevlV.  .......vPN  insSnPttSn  saPALdAaET  GhTssvqPeD  49

fa10 T.Q....VQR  Rhhtd..V.q  Fimd....rf  vki.Nslsp.  ....THvi.D  .Lmqthkhg.  60
fa12 T.Q....VQR  Rhhtd..V.S  Fimd....rf  vki.kslnp.  ....THvi.D  .Lmqthqhg.  60
folk T.Q....iQR  Rqhtd..V.S  Fimd....rf  vkV.tpqnq.  ....inil.D  .Lmqipsht.  60
   e TVks......  qS.lESgfap  .FsnGtC.pn  svilt pgpqf  dpaydqlrpq  rLteiWqngN  97
 plm TVQTRHVVQh  RSRSESsiES  FFARGACVti  mtVDNpasT.  .......tnkD  KLFaVWK.I.  110
 pls TVQTRHVVQh  RSRSESsiES  FFARGACVai  itVDNsasT.  .......knkD  KLFtVWK.I.  110
 p2s TVQTRHViQR  RtRSESsVES  FFARGACVai  ieVDNdapT.  ......krAs  rLFsVWK.I.  110
 p3s TVQTRHVVQR  RSRSEStiES  FFARGACVai  ieVDNeqpT.  ......trAq  KLFamWr.I.  108
 cb3 TmQTRHVkny  hSRSEStiEn  FlcRsACVyf  teykN.....  ......sgA.  KryaeWv.l.  92
 r14 sieTRttymh  fnqSEtdVEc  FlgRaACVhv  teiqNkdaTg  ..idnHreA.  KLFndWK.IN  105
  r2 vieTRyVqts  qtRdEmslES  FlgRsgCihe  sklevtlany  .......nke  n.FtVWa.IN  100

fa10 ...ivgal.L  Raa.......  ....TYyfsD  LEIvv.....  .......rhd  .gNLtw....  88
fa12 ...lvgal.L  Raa.......  ....TYyfsD  LEIvv.....  .......rhd  .gNLtw....  88
folk ...lvgal.L  Ras.......  ....TYyfsD  LEIav.....  .......khe  .gdLtw....  88
   e eeTs.kvfpL  ksKqdysFcl  fspfvYykcD  LEvTlsphts  q......Nhg  ...Llvrwc.  146
 plm ..TYKDtVQL  RRKLEF.F..  ....TYsRFD  mElTFVVTAN  FT..eTNNgH  AlNqvYQIMY  159
 pls ..TYKDtVQL  RRKLEF.F..  ....TYsRFD  mEfTFVVTAN  FT..eTNNgH  AlNqvYQIMY  159
 p2s ..TYKDtVQL  RRKLEF.F..  ....TYsRFD  mEfTFVVTsN  yi..DaNNgH  AlNqvYQIMY  159
 p3s ..TYKDtVQL  RRKLEF.F..  ....TYsRFD  mEfTFVVTAN  FT..naNNgH  AlNqvYQIMY  157
 cb3 ..Tprqaa QL  RRKLEF.F..  ....TYvRFD  LElTFViTst  qqPstTqNqd  AqiLthQIMY  143
 r14 L.ssl..VQL  RkKLEl.F..  ....TYvRFD  sEyTilaTAs  q.P.DsaN.y  ssNLvvQaMY  152
  r2 L...qemaQi  RRKfEl.F..  ....TYtRFD  sEITlVpcis  alsqD.....  ighitmQyMY  145

fa10 VPnGAPe...  ....aals.n  T.SN.Ptayn  kAP....ftR  lalPYtaphr  vlatv.YDG.  132
fa12 VPnGAPe...  ....aals.n  T.gN.Ptayn  kAP....ftR  lalPYtaphr  vlatv.YnG.  132
folk VPnGAPe...  ....kald.n  T.tN.Ptayh  kAP....ltR  lalPYtaphr  vlatv.YnGe  133
   e .PtGtPtkpt  tqvlhevssl  segrtPqVys  agPgisnqis  fvVPYnspls  vlSavwYnG.  204
 plm VPPGAPVPe.  ..KWDDYTWQ  TSSN.PSiFY  T..YGtAPAR  ISVPYVGisN  AYSHF.YDGF  212
 pls VPPGAPVPe.  ..KWDDYTWQ  TSSN.PSiFY  T..YGtAPAR  ISVPYVGisN  AYSHF.YDGF  212
 p2s iPPGAPiPg.  ..KWnDYTWQ  TSSN.PSVFY  T..YGapPAR  ISVPYVGiAN  AYSHF.YDGF  212
 p3s iPPGAPtPk.  ..sWDDYTWQ  TSSN.PSiFY  T..YGaAPAR  ISVPYVGlAN  AYSHF.YDGF  210
 cb3 VPPGqPVPd.  ..KvDsYvWQ  TStN.PSVFw  Te..GnAPpR  mSiPflsiqN  AYSnF.YDGw  196
 r14 VPPGAPnPk.  ..eWDDYTWQ  saSN.PSVFf  k...vgdtsR  fSVPYVGlAs  AYncF.YDGy  204
  r2 VPPGAPVP..  .nsrDDYaWQ  sgtNasvfwq  h...GqAypR  fSlPflsvAs  AYymF.YDGy  198
```

Alineamiento Múltiple de secuencias

Dynamic Programming Example

Construct an optimal of these two sequences:

G A T A C T A

G A T T A C C A

Using these scoring rules:

Match:	+1
Mismatch:	-1
Gap:	-1

DYNAMIC PROGRAMMING

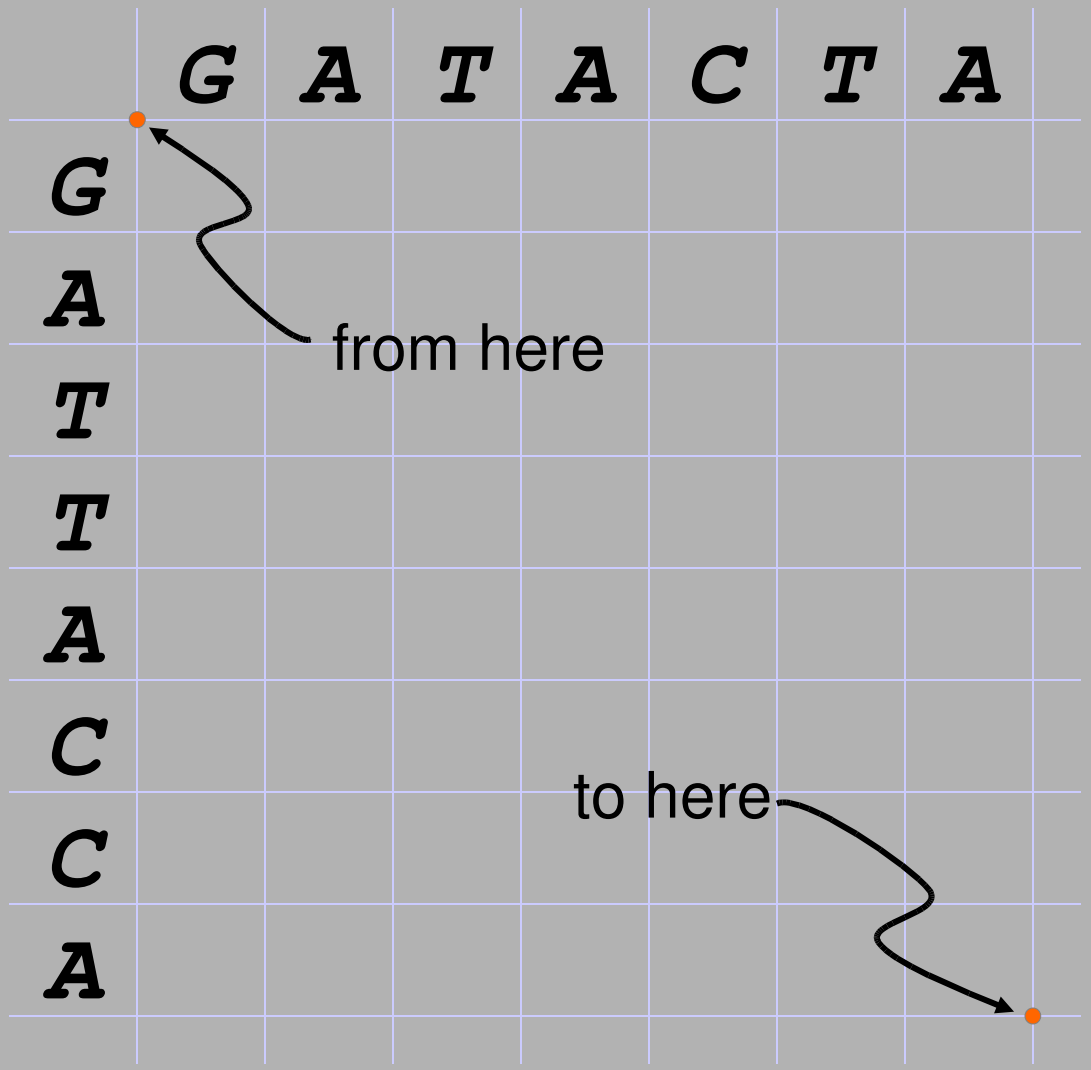
Arrange the sequence residues along a two-dimensional lattice

Vertices of the lattice fall between letters

	G	A	T	A	C	T	A	
G								
A								
T								
T								
A								
C								
C								
A								

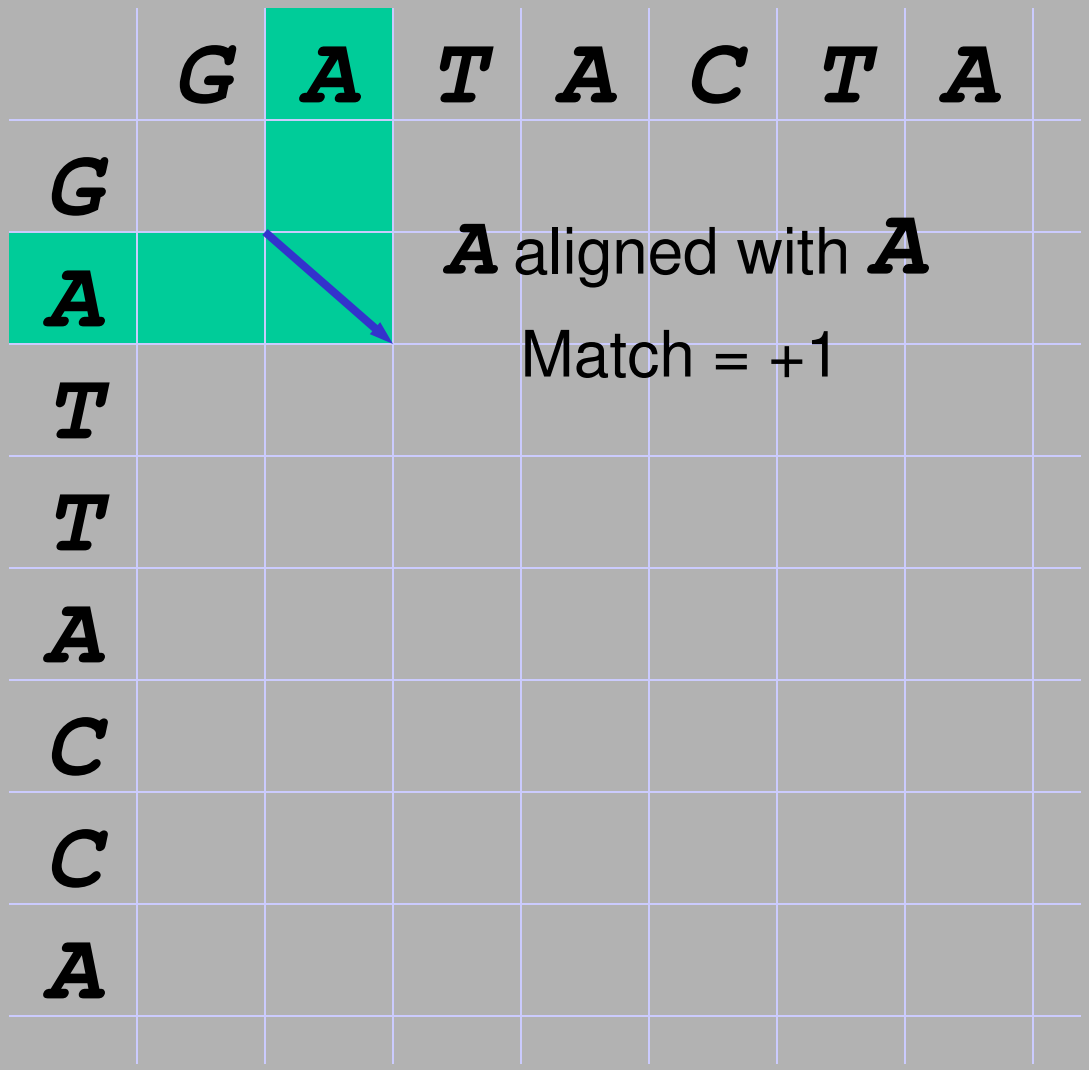
DYNAMIC PROGRAMMING

The goal is to find the optimal path



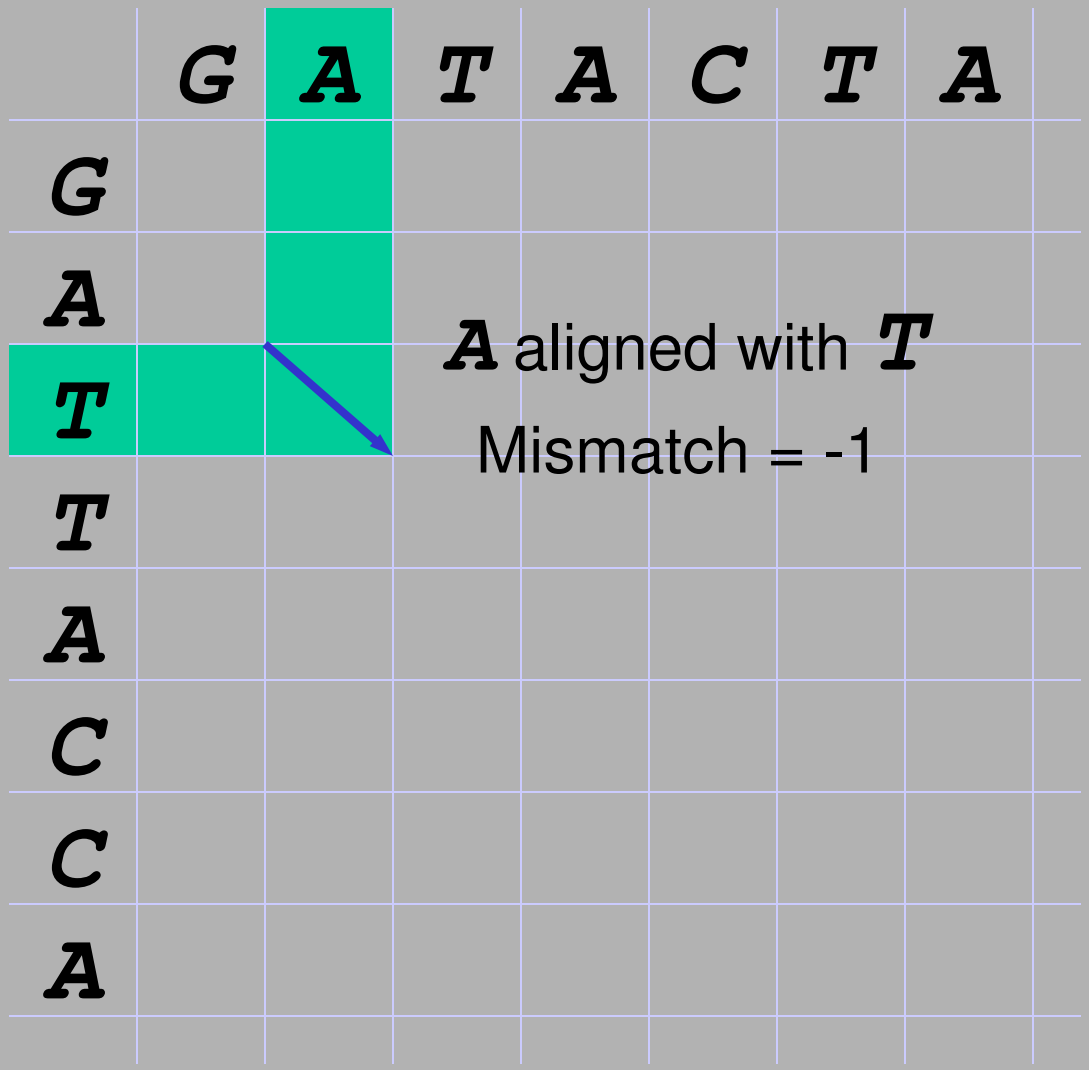
DYNAMIC PROGRAMMING

The score for a path is the sum of its incremental edges scores



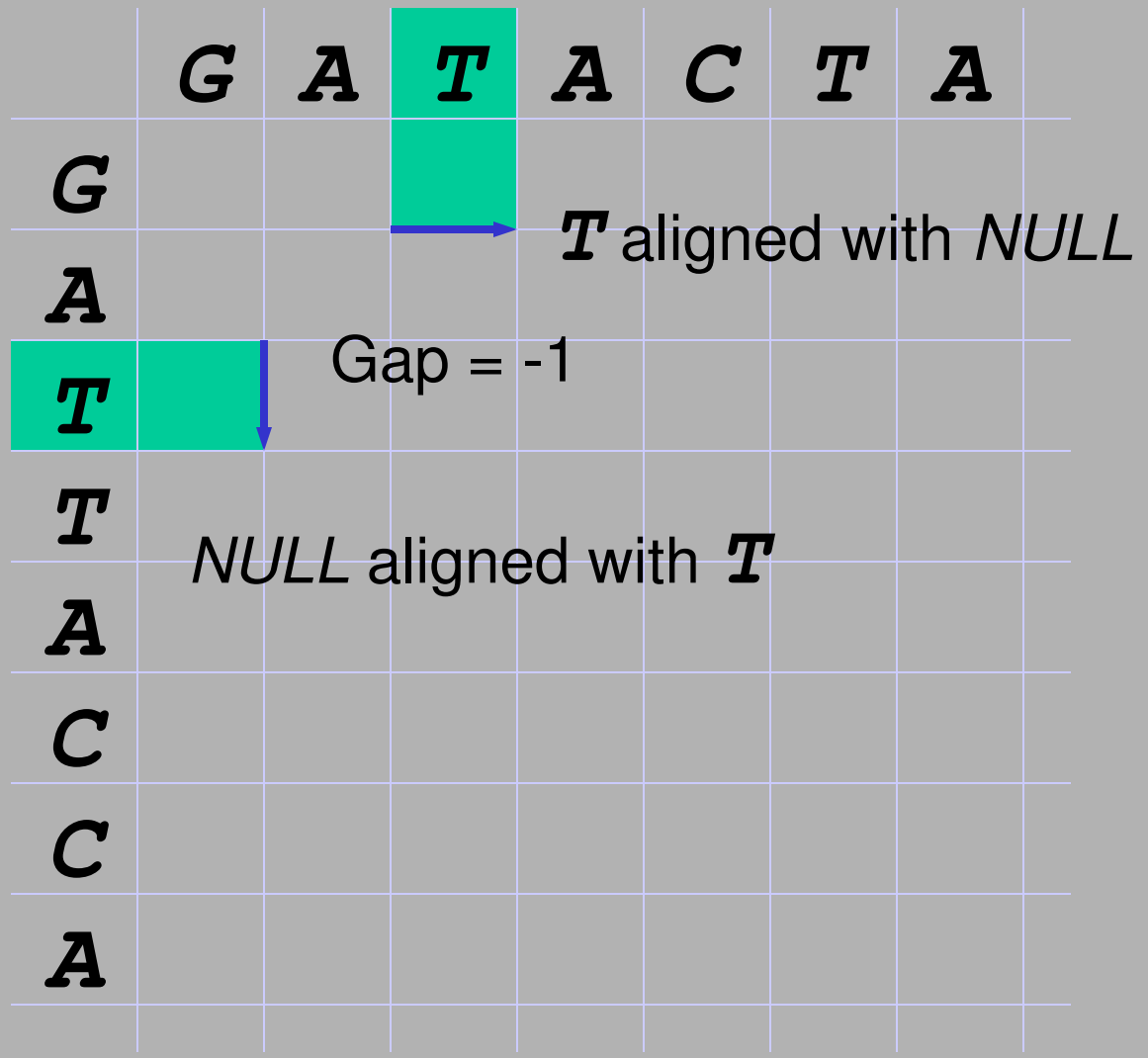
DYNAMIC PROGRAMMING

The score for a path is the sum of its incremental edges scores



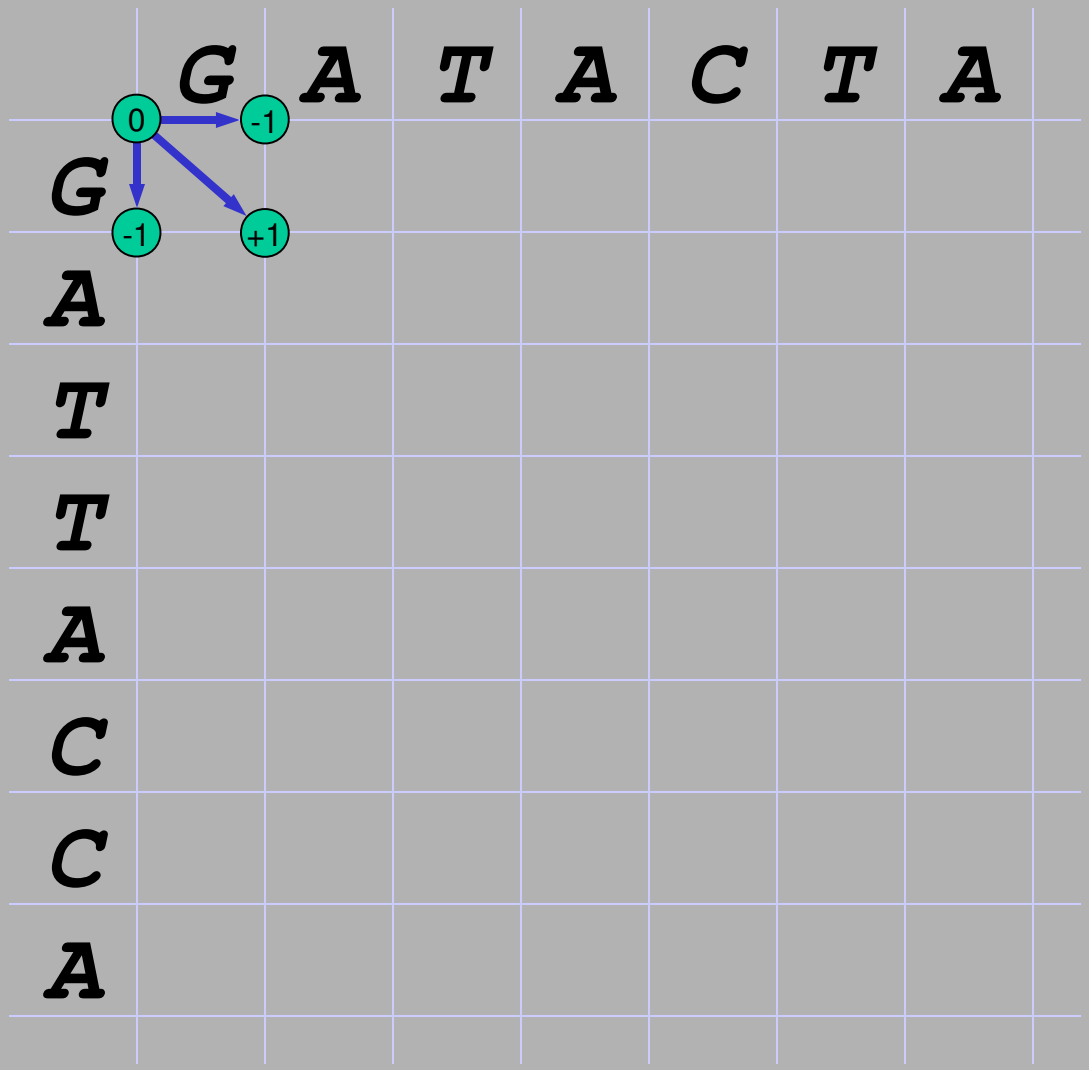
DYNAMIC PROGRAMMING

The score for a path is the sum of its incremental edges scores



DYNAMIC PROGRAMMING

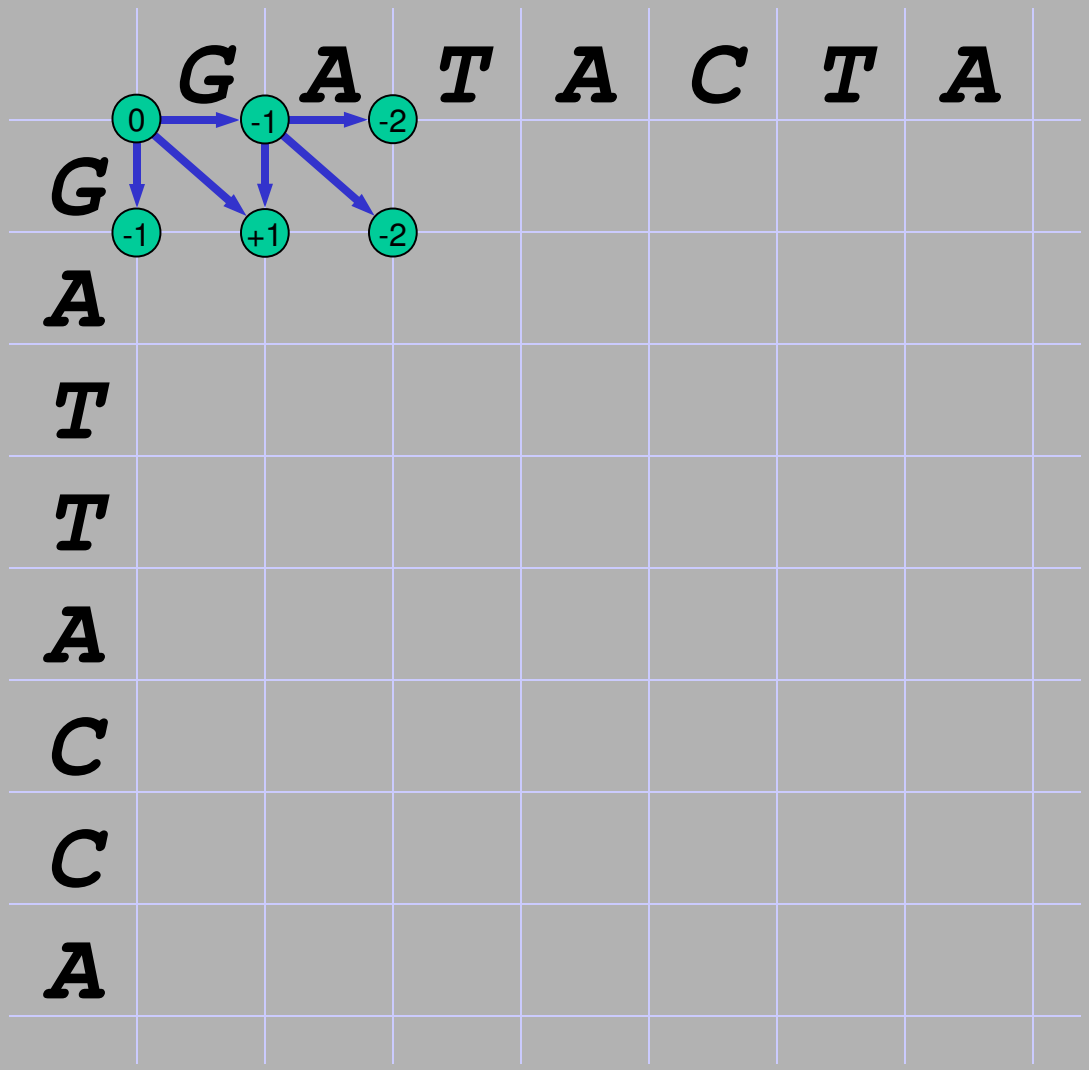
Incrementally extend the path



DYNAMIC PROGRAMMING

Incrementally extend the path

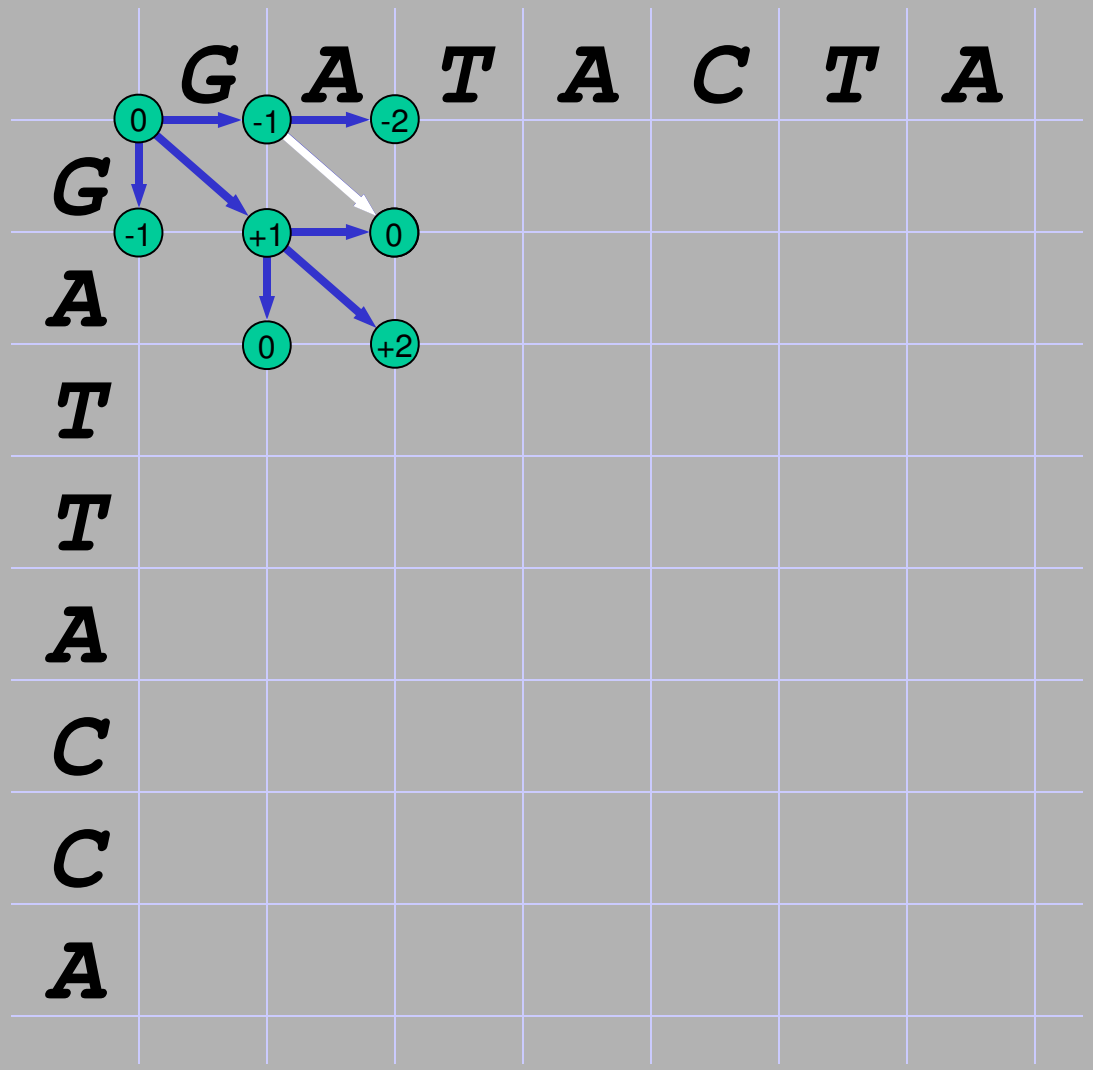
Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

Incrementally extend the path

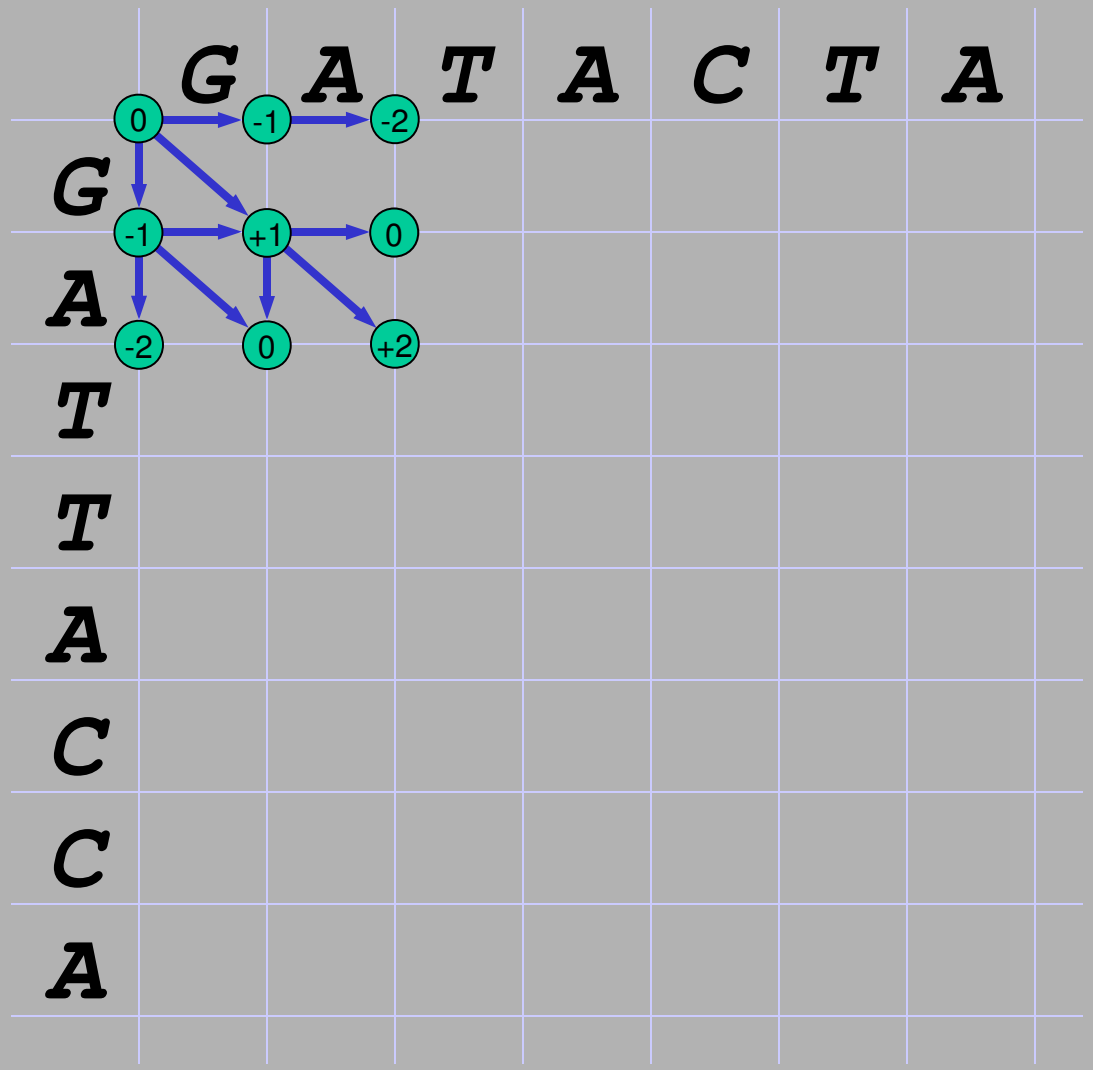
Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

Incrementally extend the path

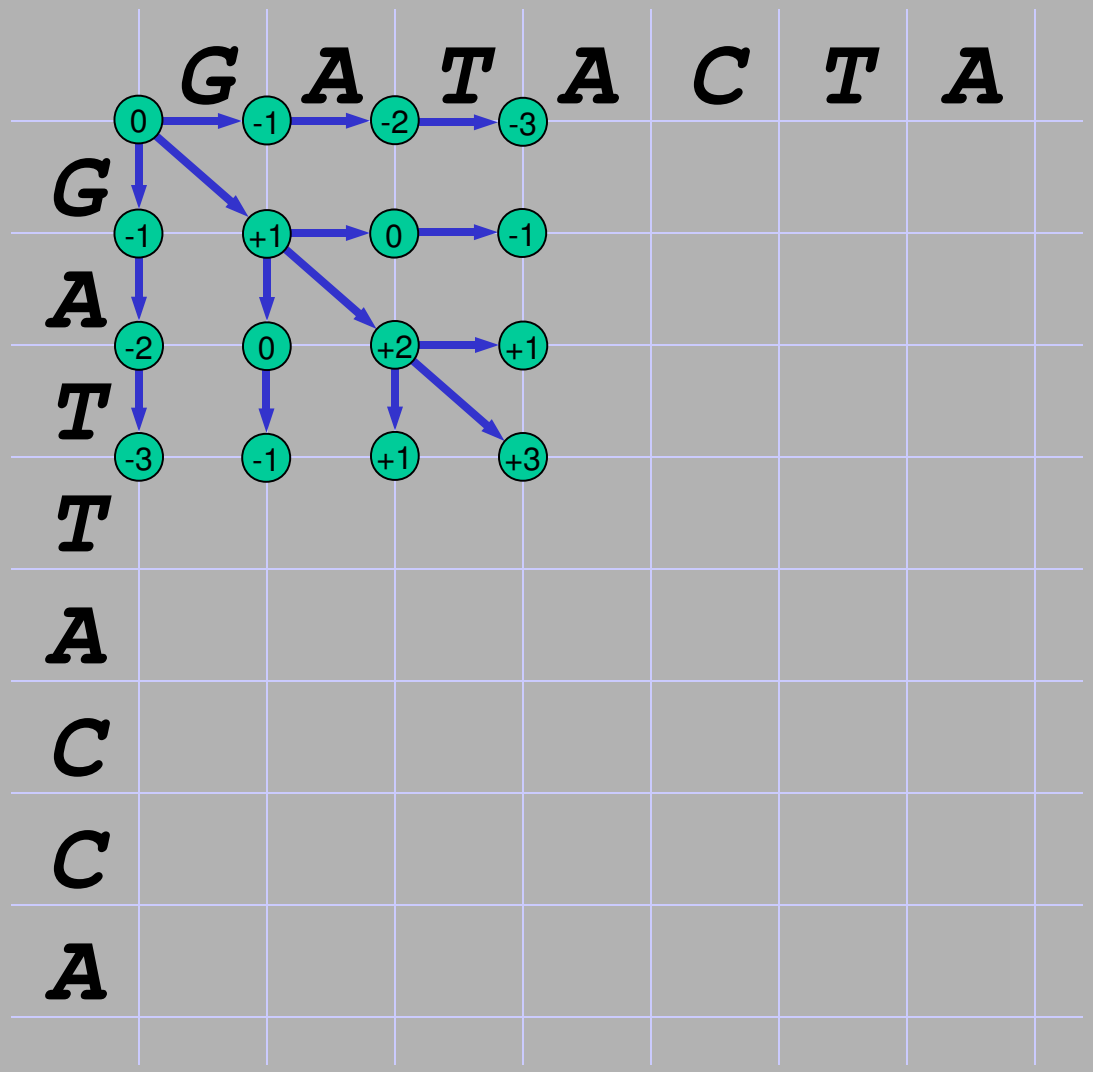
Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

Incrementally extend the path

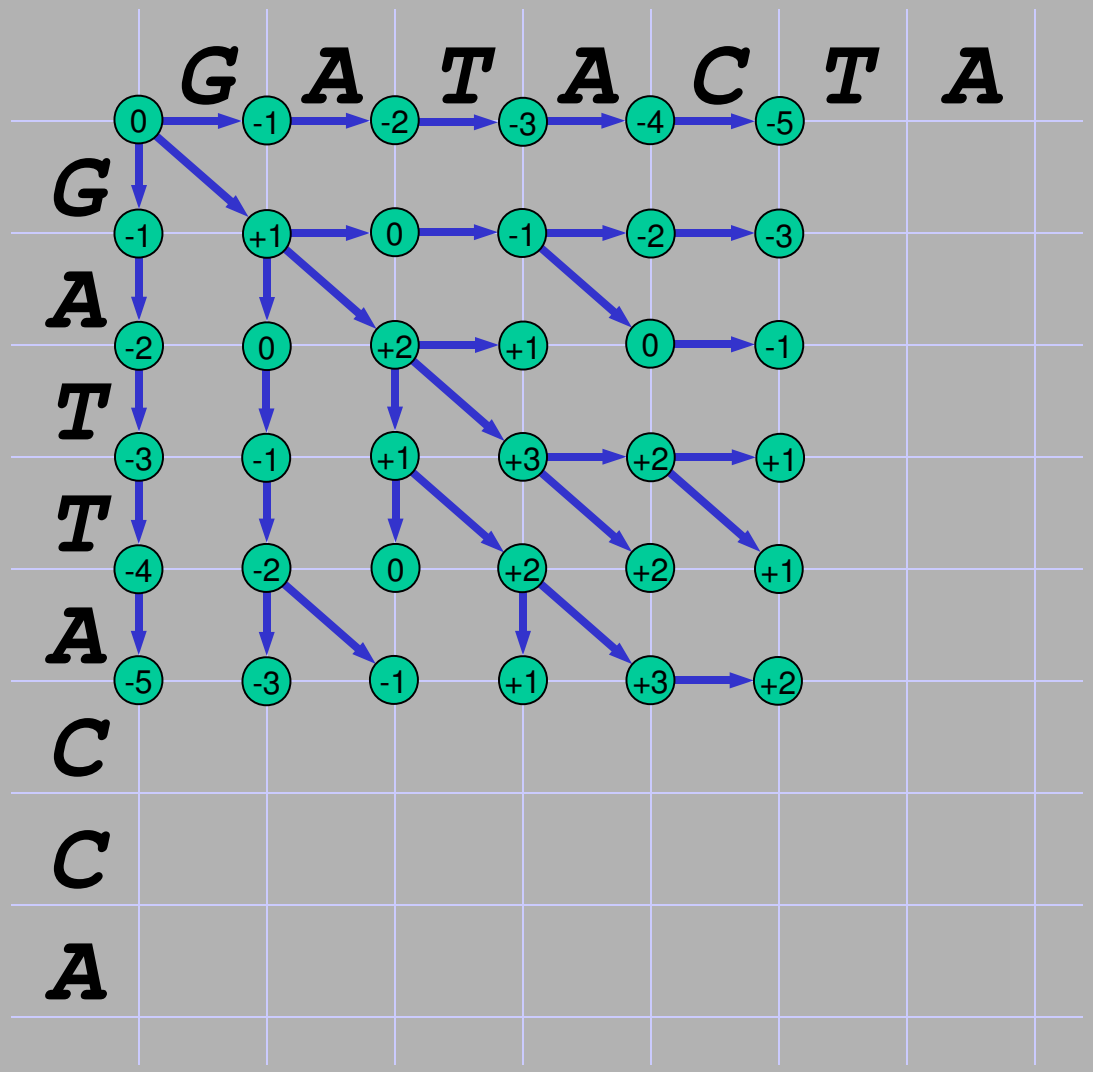
Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

Incrementally extend the path

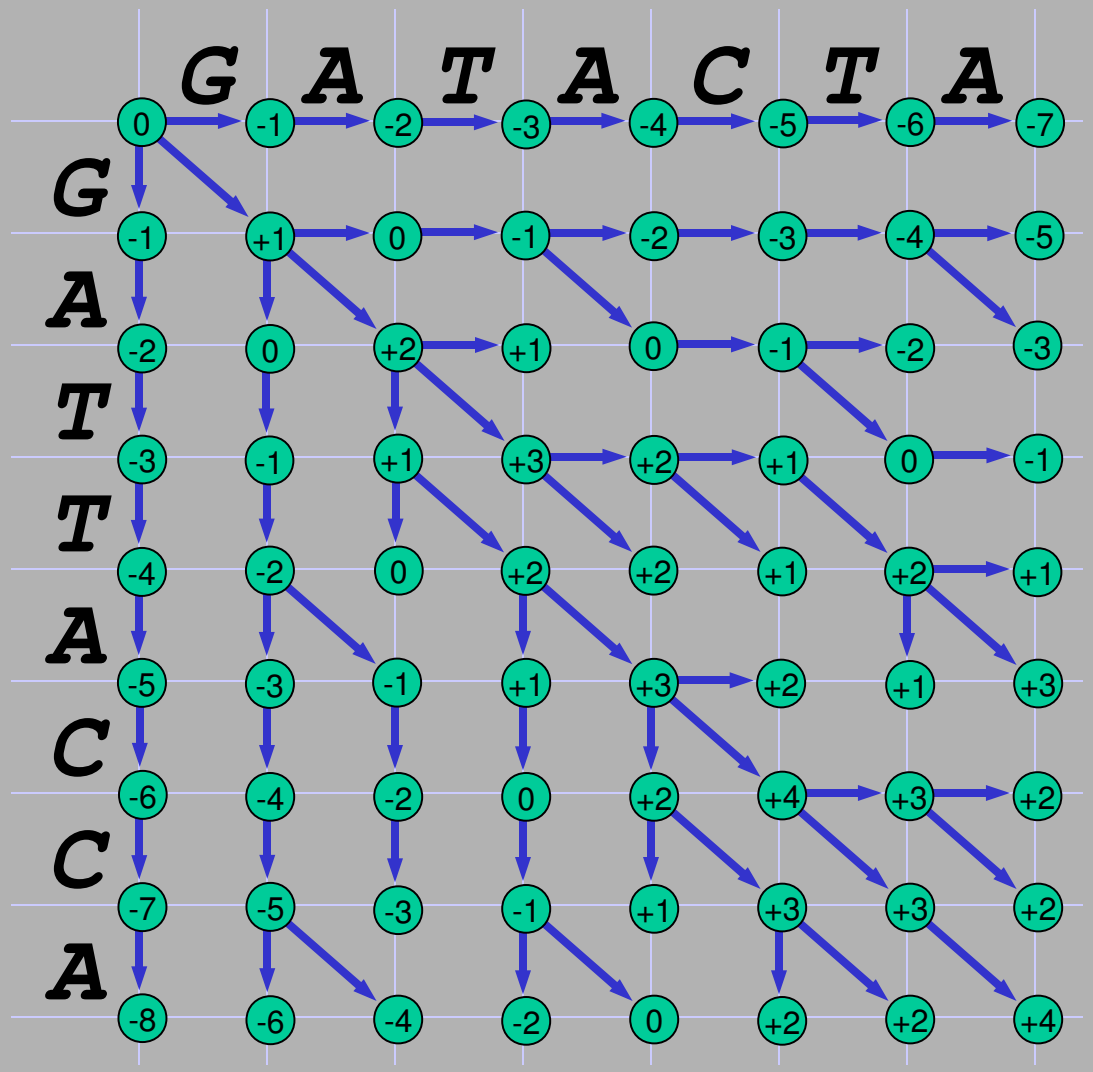
Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

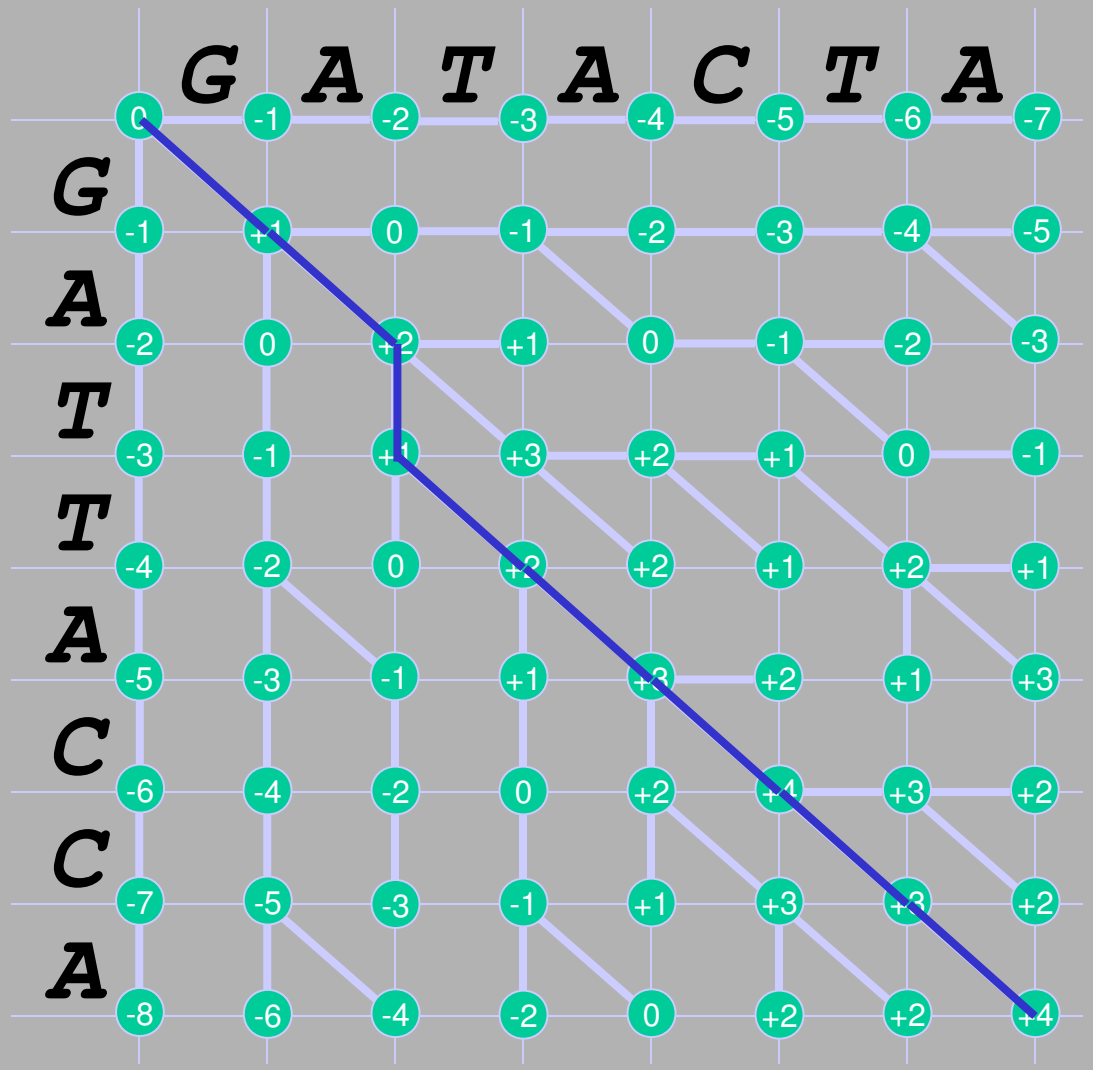
Incrementally extend the path

Remember the best sub-path leading to each point on the lattice



DYNAMIC PROGRAMMING

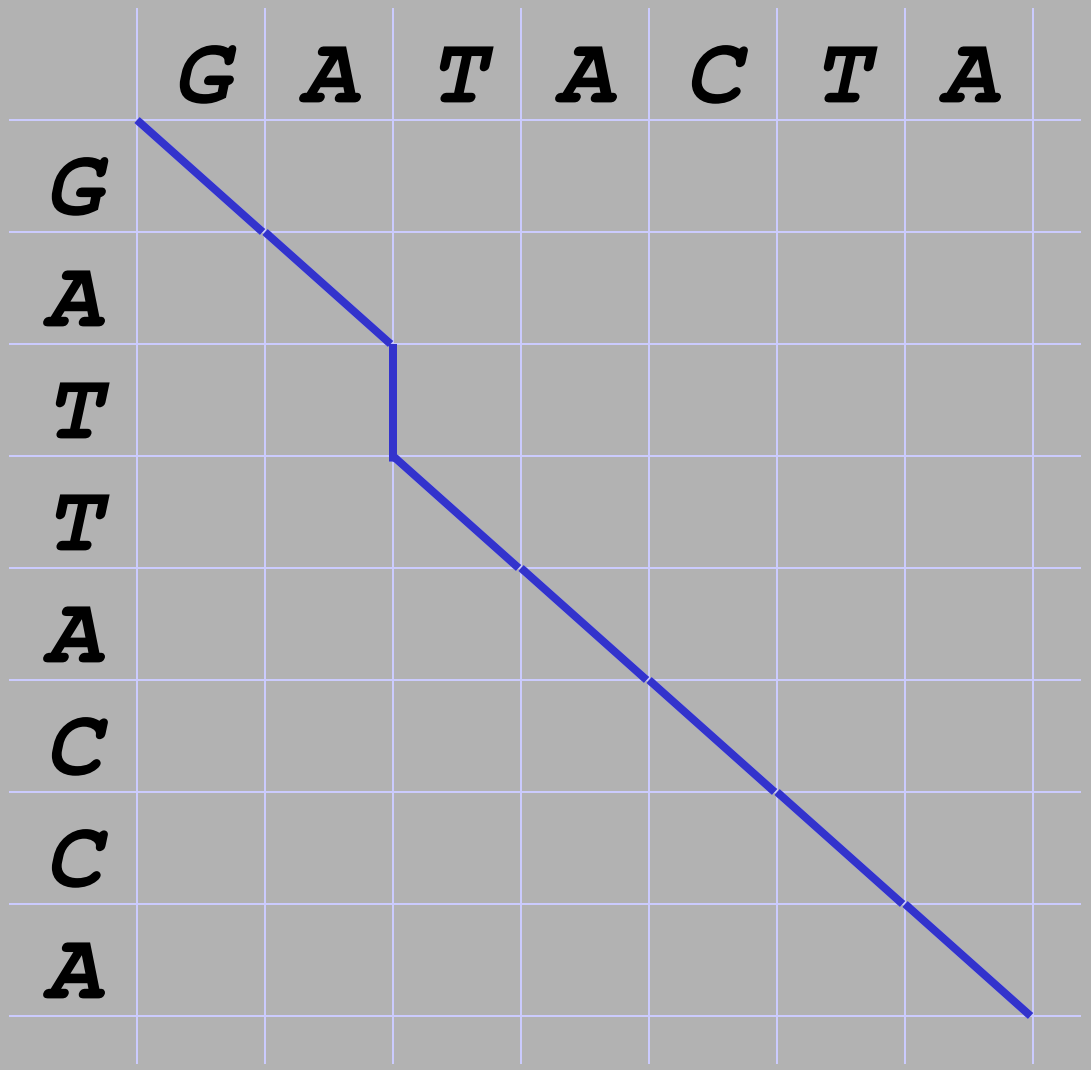
Trace-back to get optimal path and alignment



DYNAMIC PROGRAMMING

Print out the alignment

GA-TACTA
GATTACCA



HOW CAN I ALIGN *MANY* SEQUENCES

2 Globins =>1 Min

HOW CAN I ALIGN *MANY* SEQUENCES

3 Globins =>2 hours

HOW CAN I ALIGN *MANY* SEQUENCES

4 Globins => 10 days

HOW CAN I ALIGN *MANY* SEQUENCES

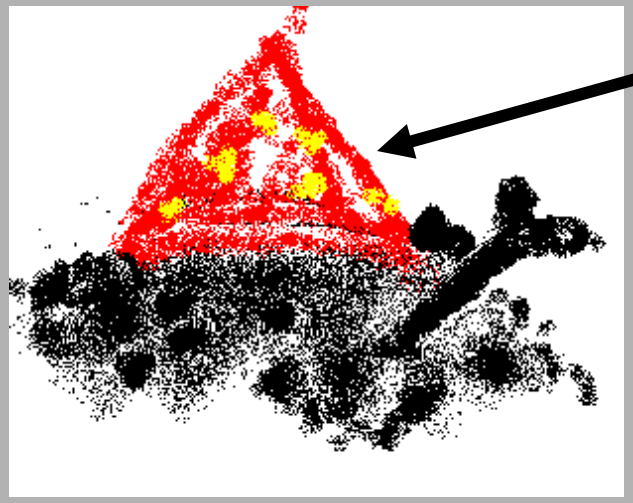
5 Globins => 3 years

HOW CAN I ALIGN *MANY* SEQUENCES

6 Globins =>300 years

HOW CAN I ALIGN *MANY* SEQUENCES

7 Globins =>30. 000 years



Solidified Fossil, Old stuff

HOW CAN I ALIGN *MANY* SEQUENCES

8 Globins =>3 Million years

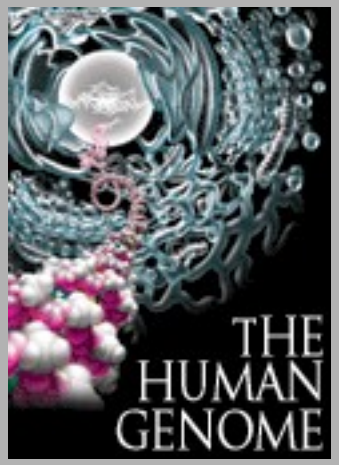
Commercial Structural Genomics Initiatives

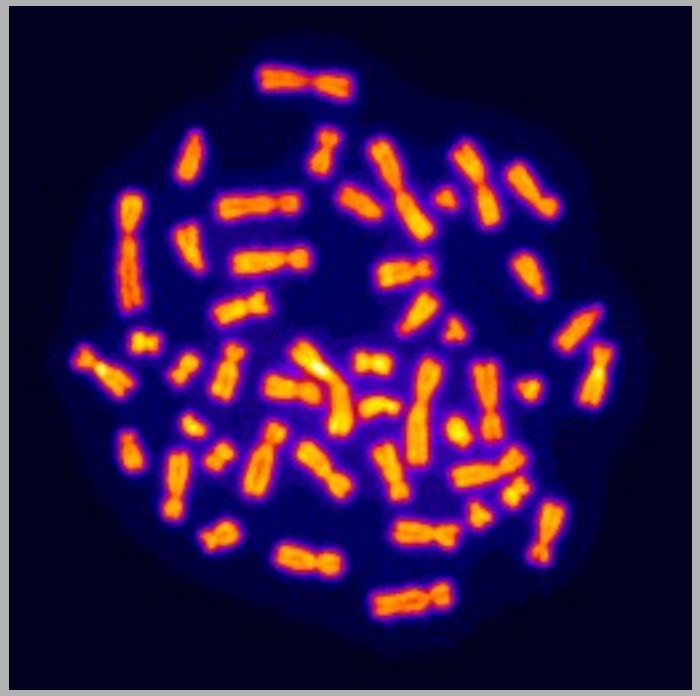
- IBM (Blue Gene project: 2000)
  - Computational protein folding
- Geneformatics (1999)
  - Modeling for identifying active sites
- Prospect Genomics (1999)
  - Homology modeling
- Protein Pathways (1999)
  - Phylogenetic profiling, domain analysis, expression profiling
- Structural Bioinformatics Inc (1996)
  - Homology modeling, docking

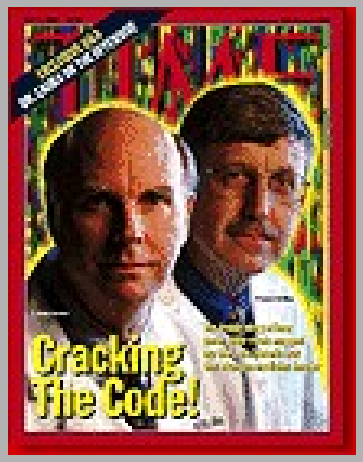
Proyecto Genoma Humano

La secuencia del genoma está casi completa!

- aproximadamente 3.5 billones de pares de bases.







Raw Genome Data

Implications for Biomedicine

- Physicians will use genetic information to diagnose and treat disease.
  - Virtually all medical conditions (other than trauma) have a genetic component.
- Faster drug development research
  - Individualized drugs
  - Gene therapy
- All Biologists will use gene sequence information in their daily work

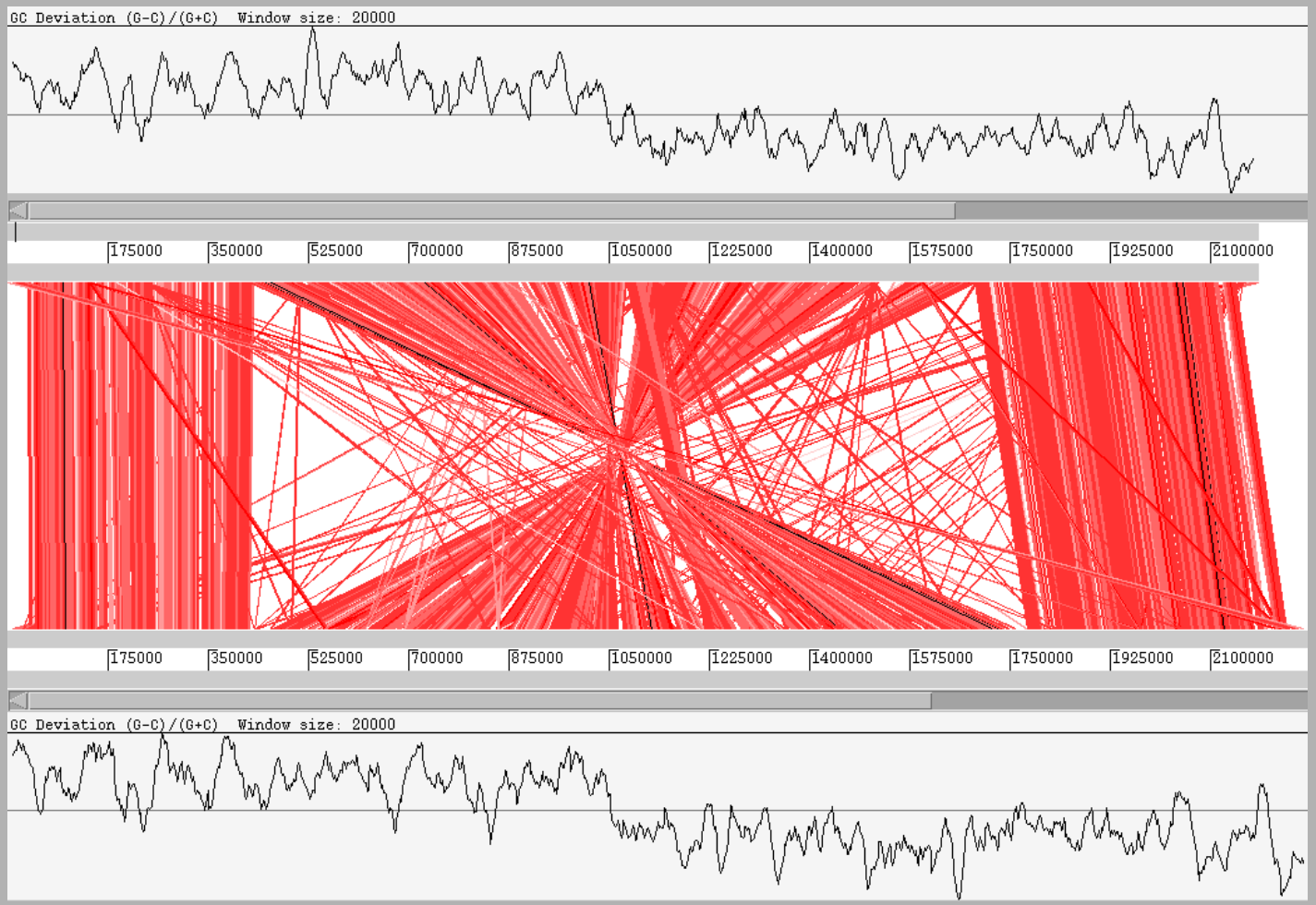
Bioinformatics Challenges

The huge dataset

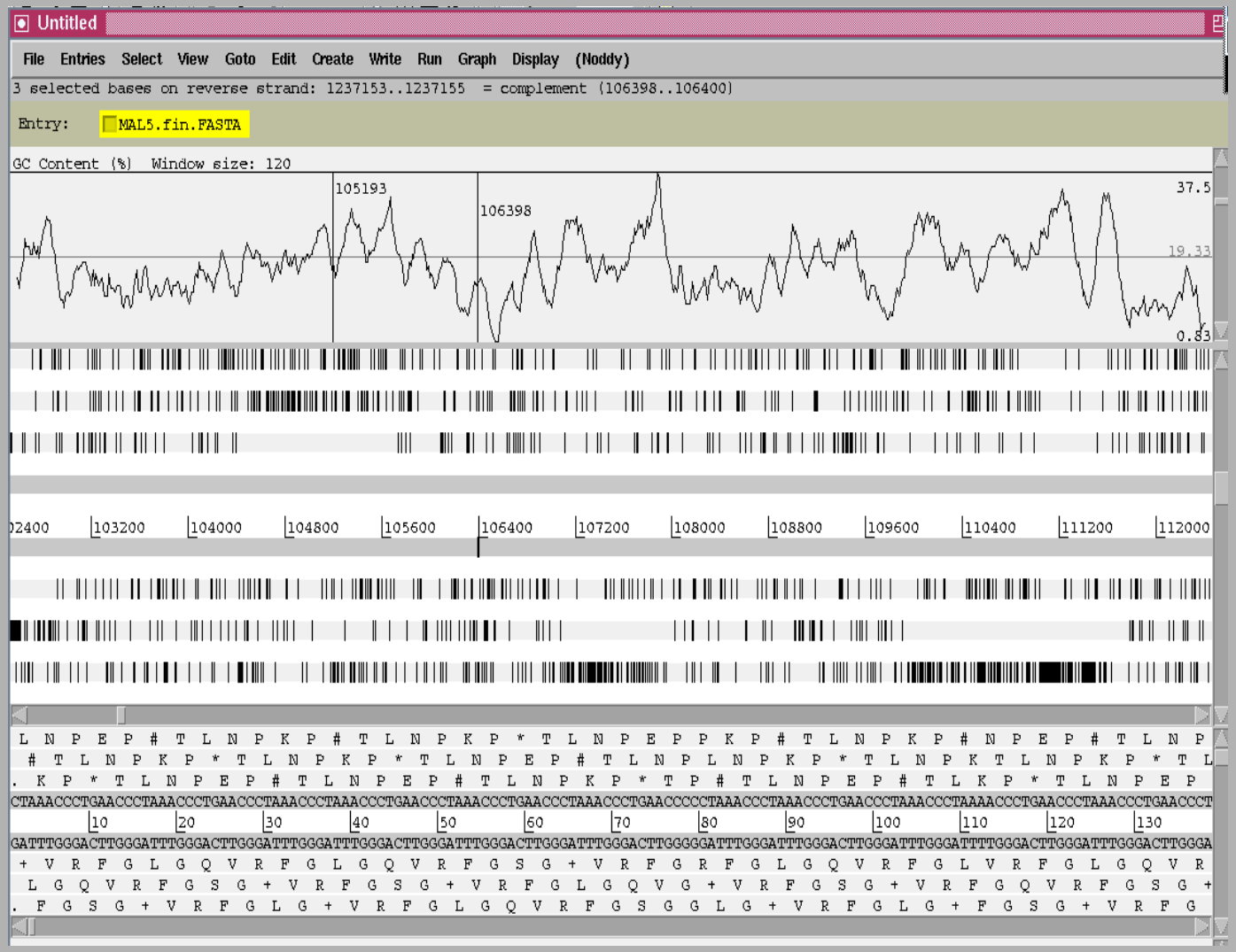
- Lots of new sequences being added
  - automated sequencers
  - Human Genome Project
  - EST sequencing
- GenBank has over 10 Billion bases and is doubling every year!!
  (problem of exponential growth...)
- How can computers keep up?

Genome comparisons

- Designed for looking at complete bacterial genomes.



Gene finding



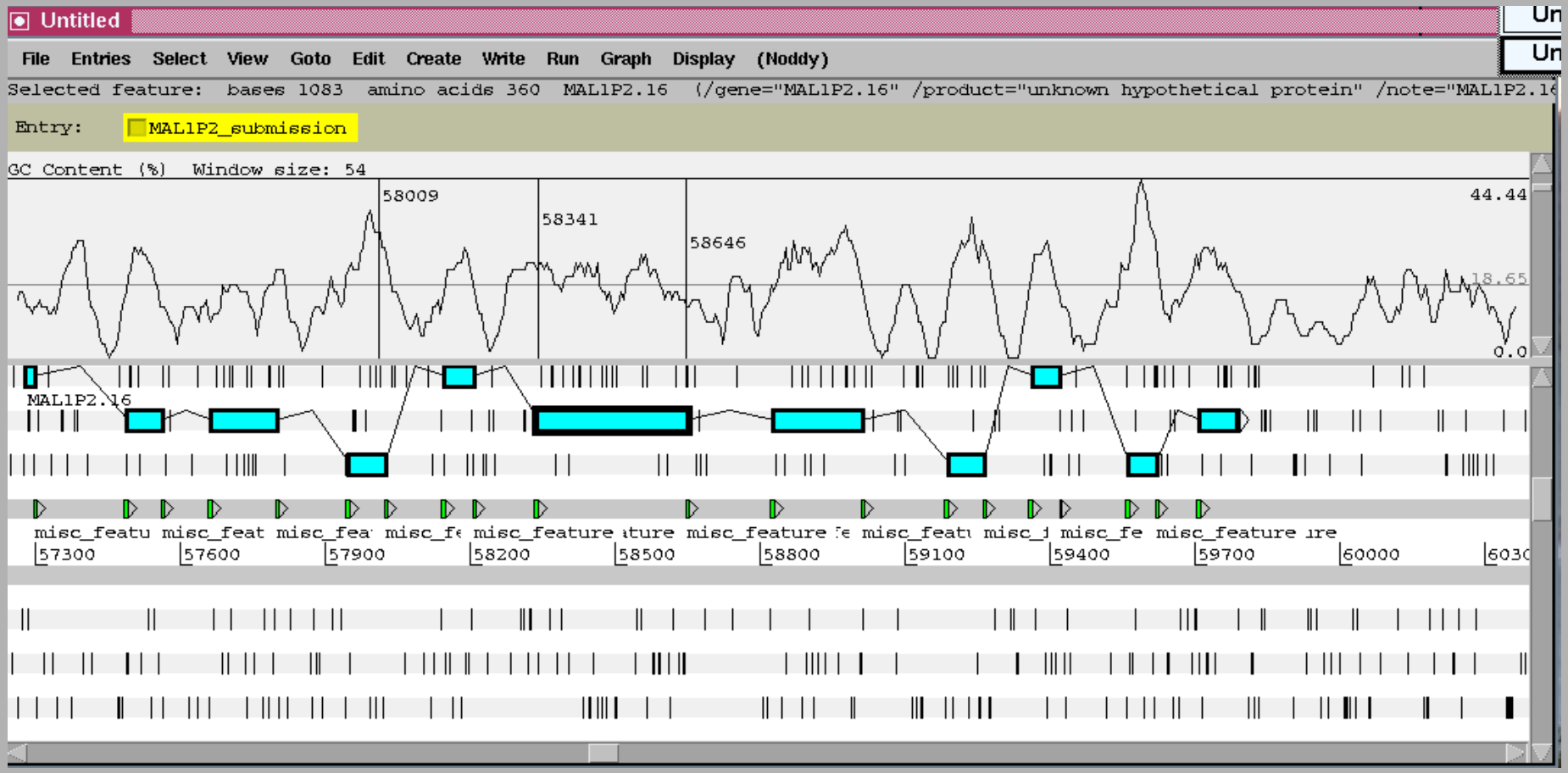
AT content

Forward translations

Reverse Translations

DNA and amino acids

Gene finding



Bringing a New Drug to Market

Review and approval by Food & Drug Administration

1 compound approved

Phase III: Confirms effectiveness and monitors adverse reactions from long-term use in 1,000 to 5,000 patient volunteers.

Phase II: Assesses effectiveness and looks for side effects in 100 to 500 patient volunteers.

Phase I: Evaluates safety and dosage in 20 to 100 healthy human volunteers.

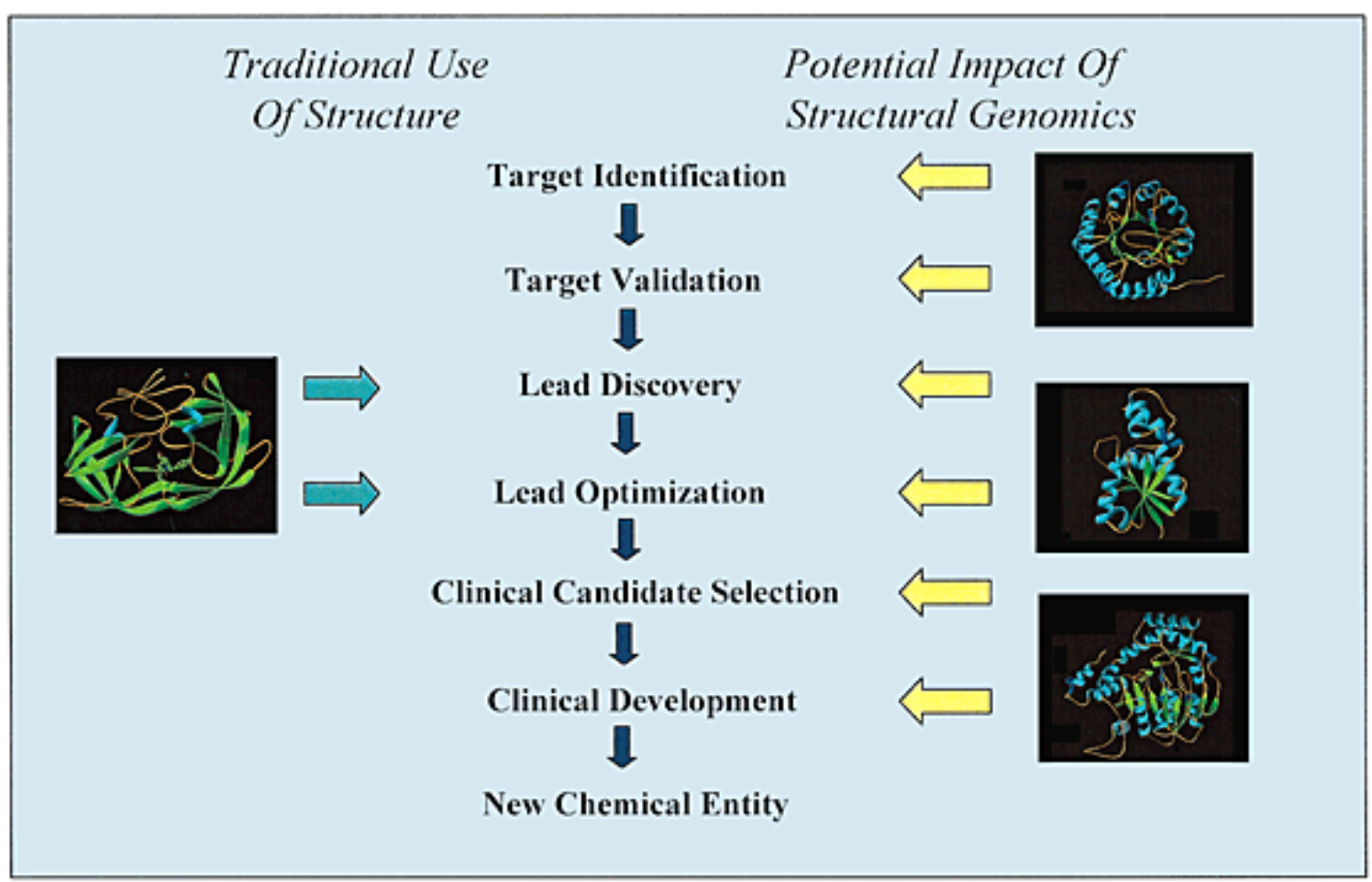
5 compounds enter clinical trials

5,000 compounds evaluated

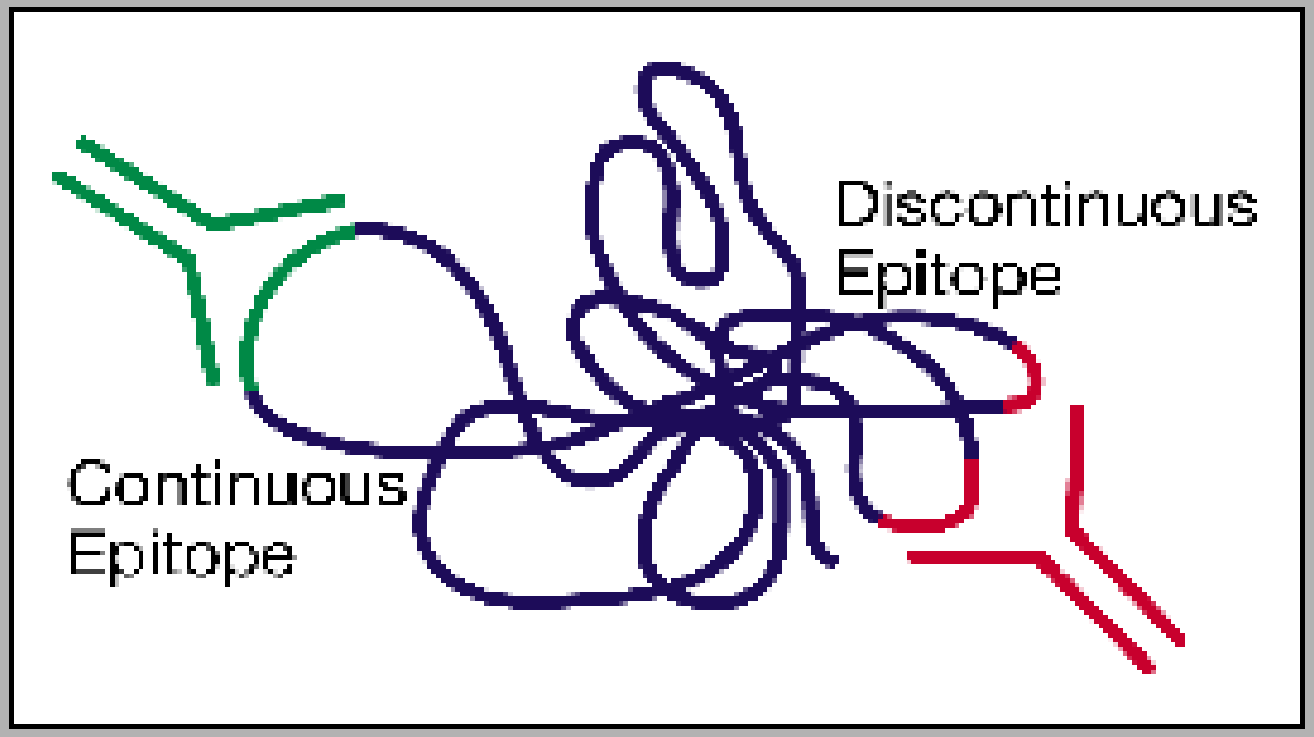
Discovery and preclininal testing: Compounds are identified and evaluated in laboratory and animal studies for safety, biological activity, and formulation.

0 2 4 6 8 10 12 14 Years 16

Impact of Structural Genomics on Drug Discovery

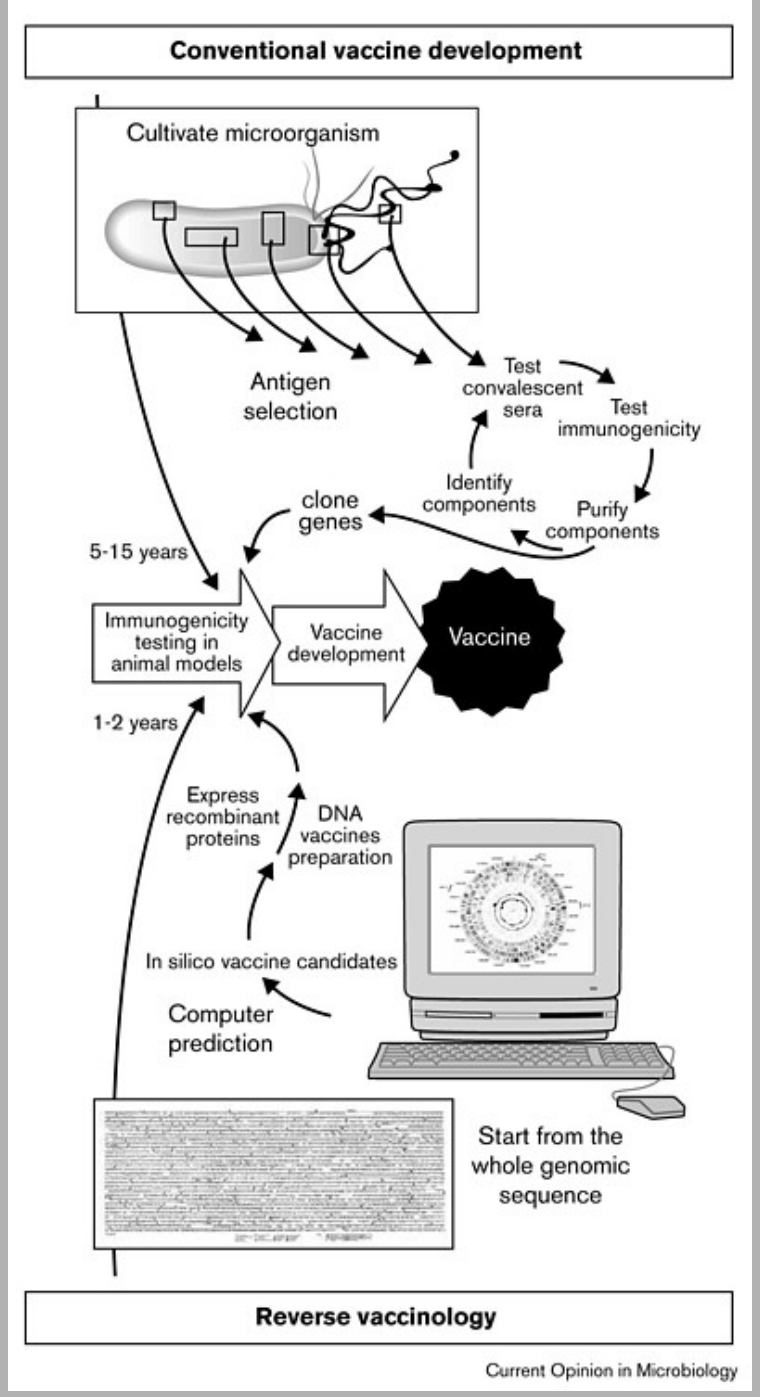


Epitopes …



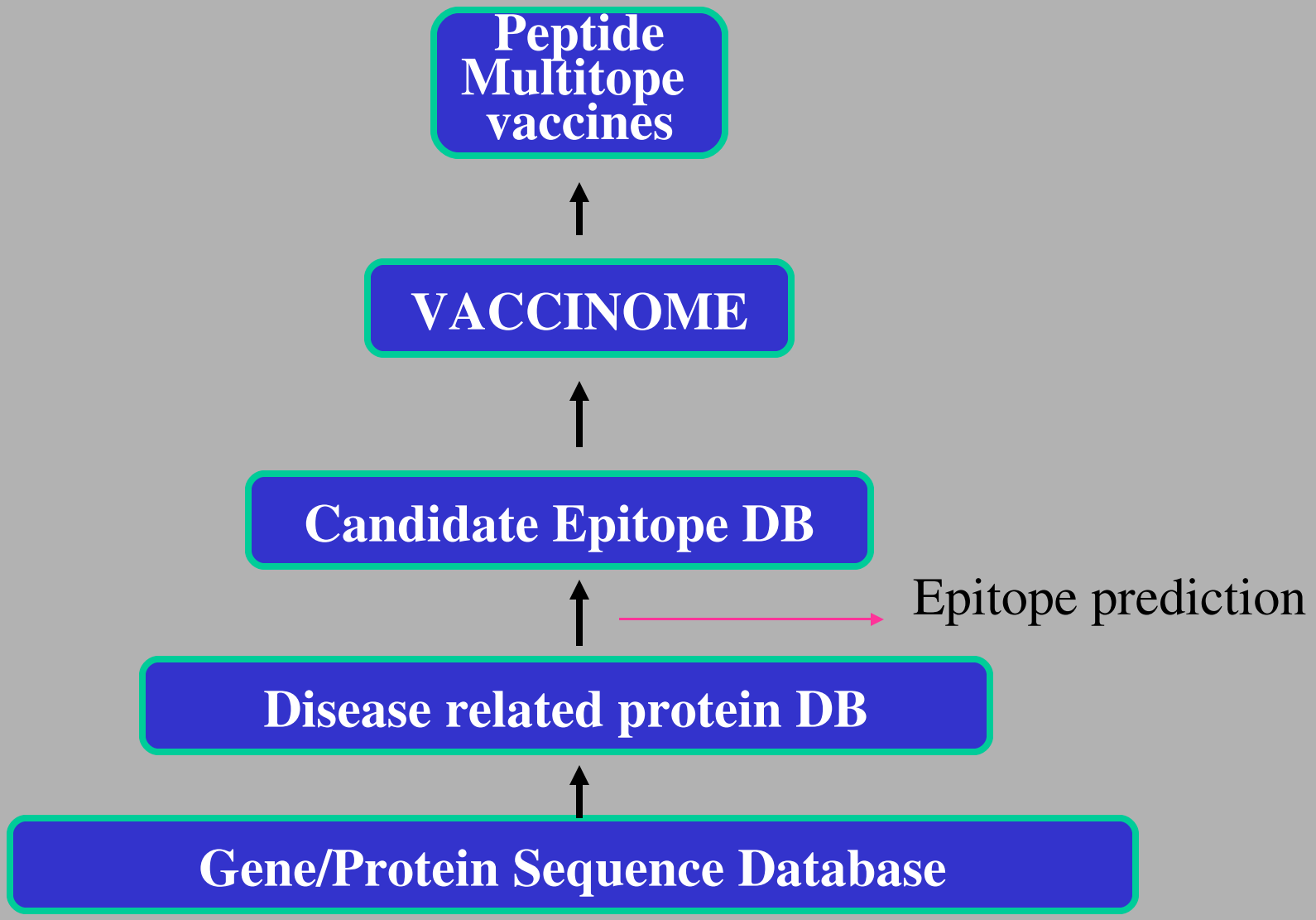
B-cell epitopes

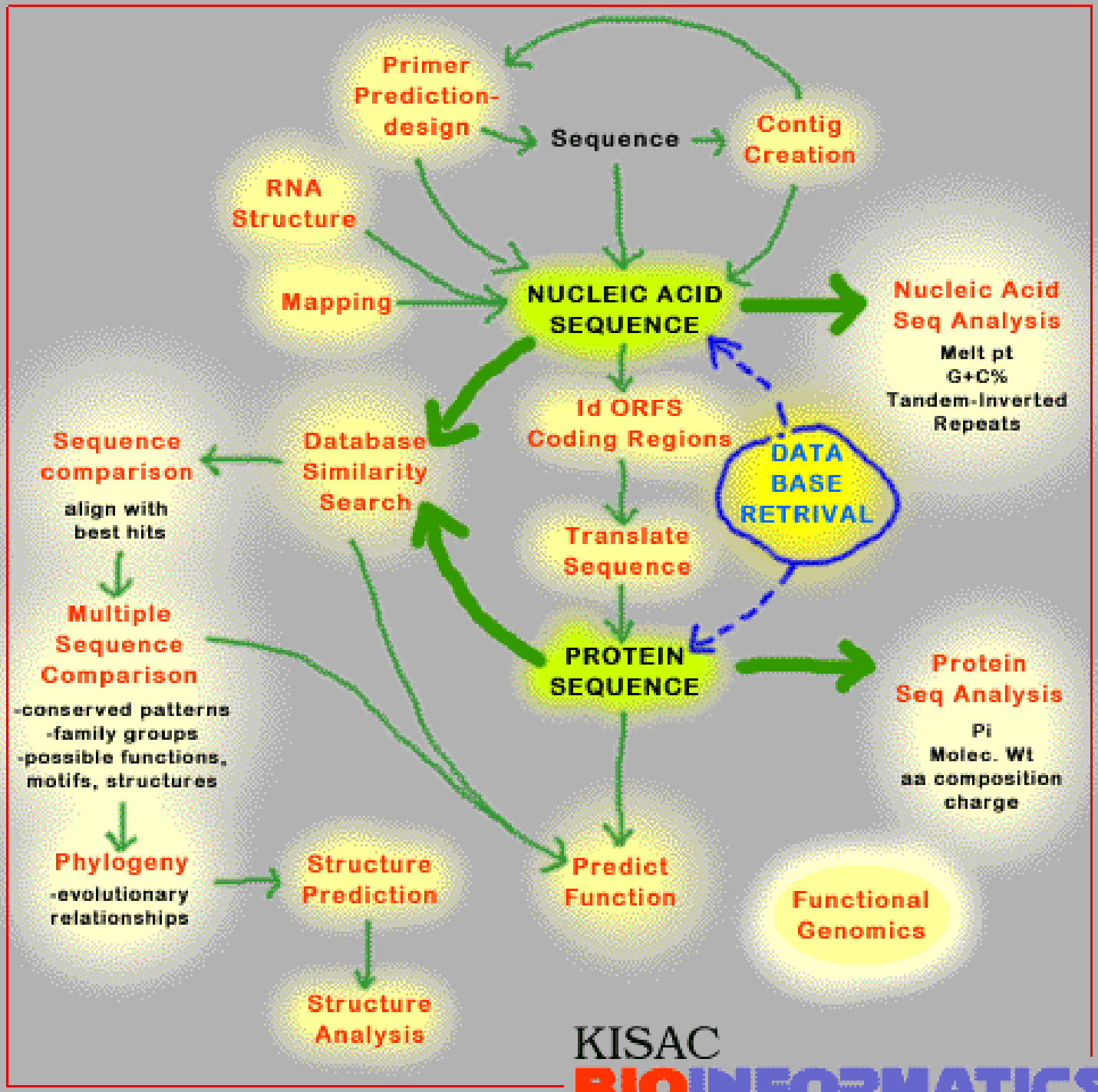
T_h-cell epitopes



Vaccine development In Post-genomic era: Reverse Vaccinology Approach.

In Silico Analysis




Primer Prediction-design
Sequence
Contig Creation
RNA Structure
Mapping
NUCLEIC ACID SEQUENCE
Nucleic Acid Seq Analysis
Melt pt
G+C%
Tandem-Inverted Repeats
Id ORFS Coding Regions
DATA BASE RETRIVAL
Sequence comparison
align with best hits
Database Similarity Search
Translate Sequence
Multiple Sequence Comparison
-conserved patterns
-family groups
-possible functions, motifs, structures
PROTEIN SEQUENCE
Protein Seq Analysis
Pi
Molec. Wt
aa composition
charge
Phylogeny
-evolutionary relationships
Structure Prediction
Predict Function
Functional Genomics
Structure Analysis
KISAC
BIOINFORMATICS

Biological Research in 21st Century

“ The new paradigm, now emerging is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical.”

- Walter Gilbert

The job of the biologist is changing

- As more biological information becomes available …
  - The biologist will spend more time using computers
  - The biologist will spend more time on data analysis (and less doing lab biochemistry)
  - Biology will become a more quantitative science (think how the periodic table and atomic theory affected chemistry)

Un ejemplo …

Cisteíno proteasa de la fasciola hepática: En busca de un péptido inmunogénico

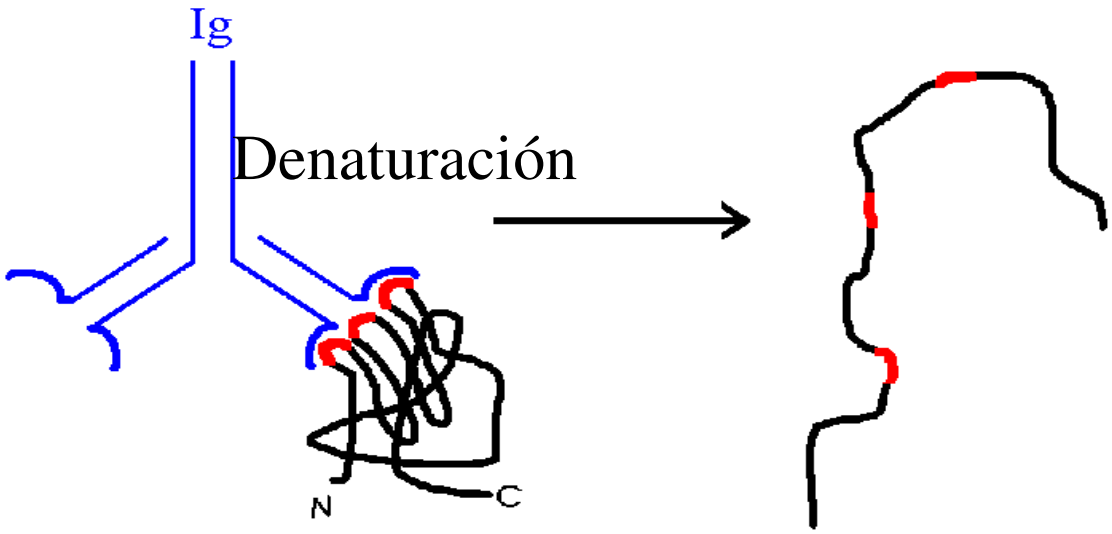
Alineamiento: cisteíno proteasas de mamífero Vs. cisteíno proteasa de *Fasciola hepatica*.

AA Idénticos

AA divergentes

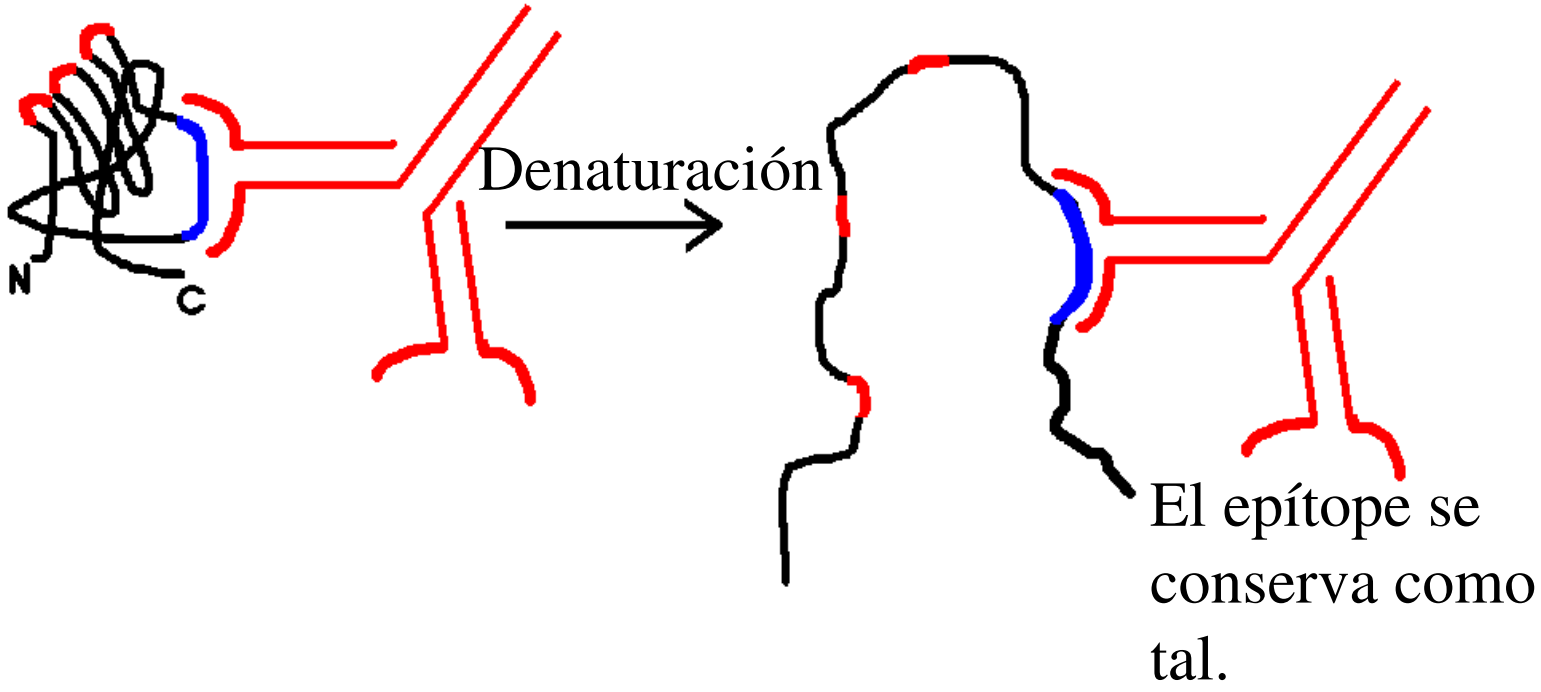
```
 VPKSVDWREKGYVTPVKNQGQCGSCWAFSATGALEGQMFRKTGR  ISLSEQNLVDCSRPQGN
AVPDKIDWRESGYVTEVKDQGNCGSCWAFSTTGTMEGQYM KNERTSISFSEQQLVDCSRPWGN
                                        _____ROJO__________

QGCNGGLMDNAFQYIKENGGLDSEESYPYEATDTSCNY KPEYSVANDTGFVDIPQREKA  LMK
NGCGGGLMENAYQYLKQF GLETESSYPYTAVGGQCRYNKQLG VAKVTGYYTV QSGSEVEL KN
                                        _VIOLETA____  _AMARILLO_______

AVATVGPISVAIDAGHSFQFYKSGIYYEPDCSSKDLDHGVLVVGYGFEG TDSNNNKYW IVKNSW
LIGSEGPSAVAVDVESDFMMYRSGIYQSQTCSPLRVNHAVLAVGYGTQGGTD     YW IVKNSW
_____                       _VERDE_____

GPEWGM-GYVKMAKDRNNH  CGIATAASYPTV
GLSWGERGYIRMV  RNRGNMCGIASLASLPMVARFP
```

Epítope Discontinuo, formado por porciones distantes de la secuencia.

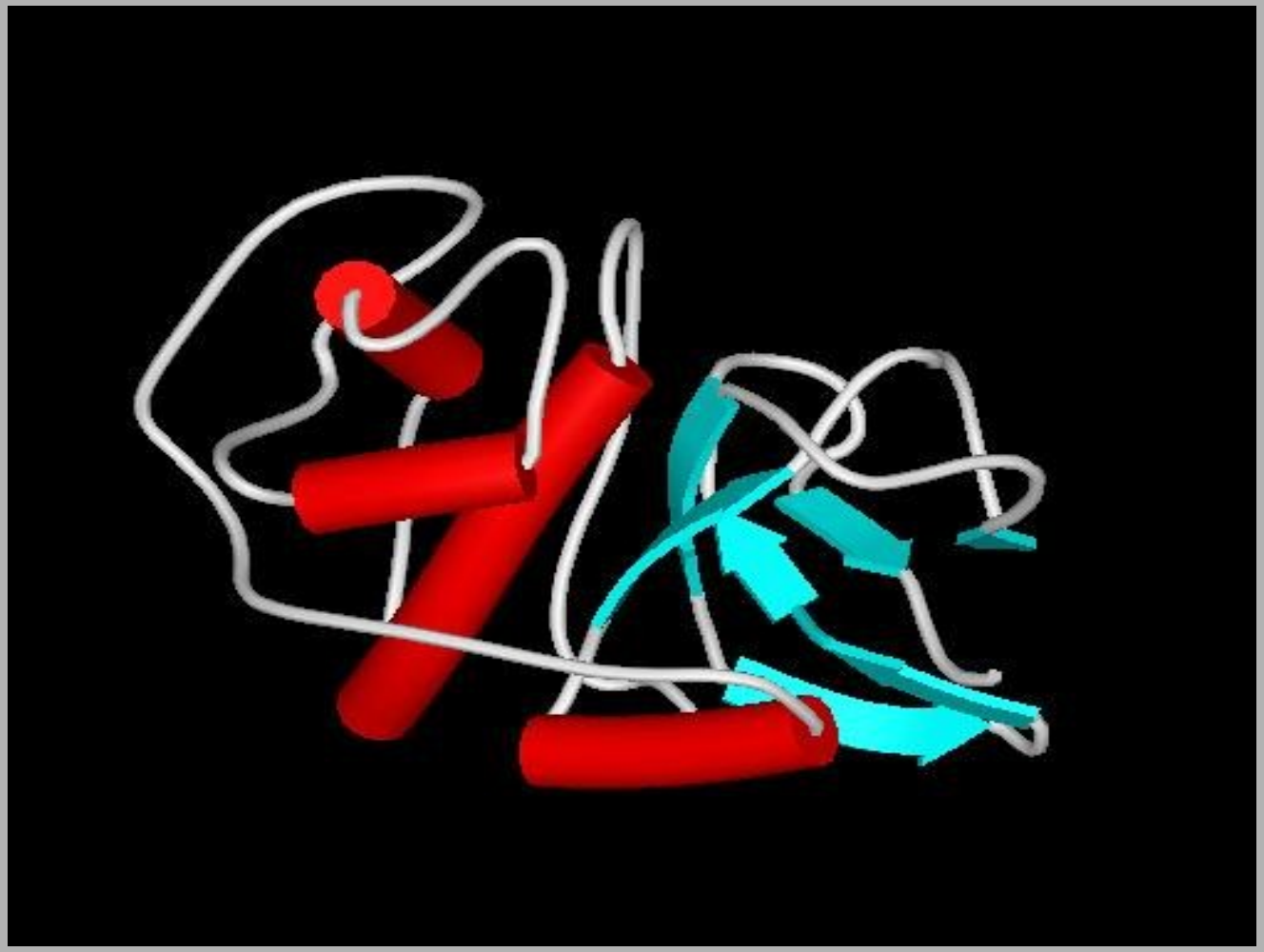


El epítope se pierde con la denaturación.

Epítope Continuo, formado por una porción de la secuencia



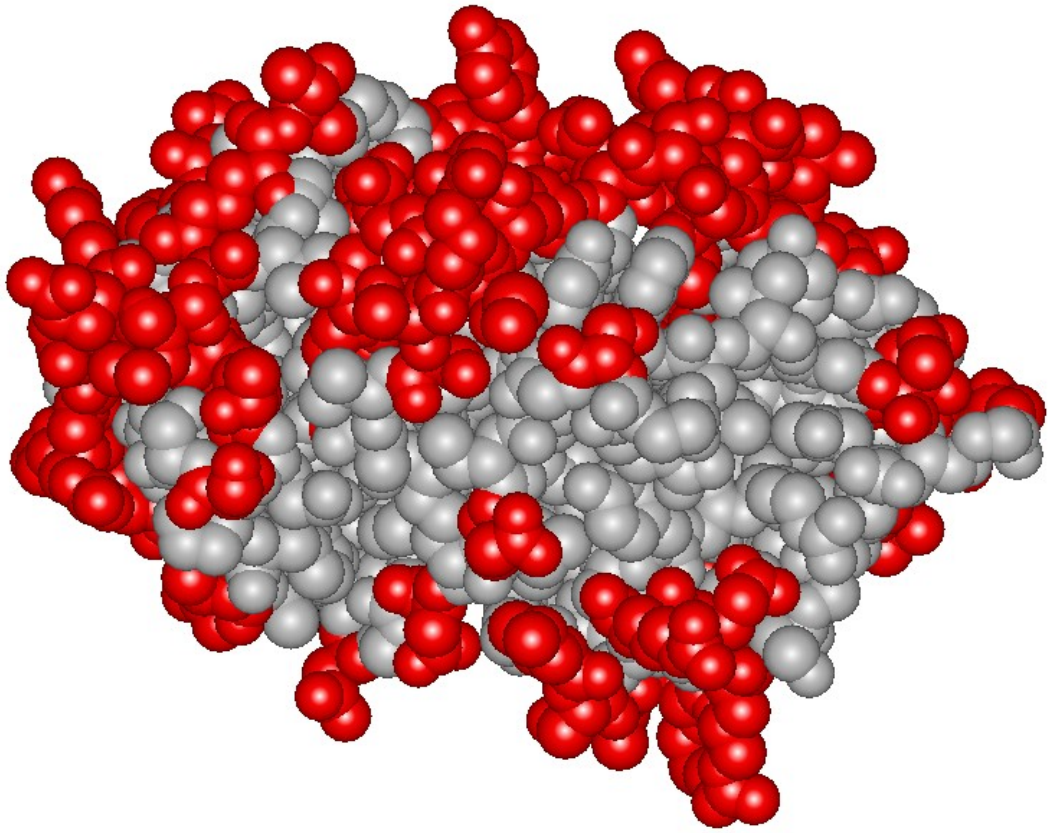
Modelaje tridimensional por homología. Identidad de secuencia de 56% con quimopapaína (1YAL)



Análisis de Superficie: vista de átomos por radio de van der Waals

AA idénticos

AA divergentes



Selección de secuencias (1)divergentes, (2)accesibles al solvente y (3)contínuas.

TMEGQYMKNERTSISFS

YYTVQSGSEVELK
NLIGSE

QSQTCSPLRVN

RYNKQLGVAKV