Introduction to Bioinformatics

D.Rabiei

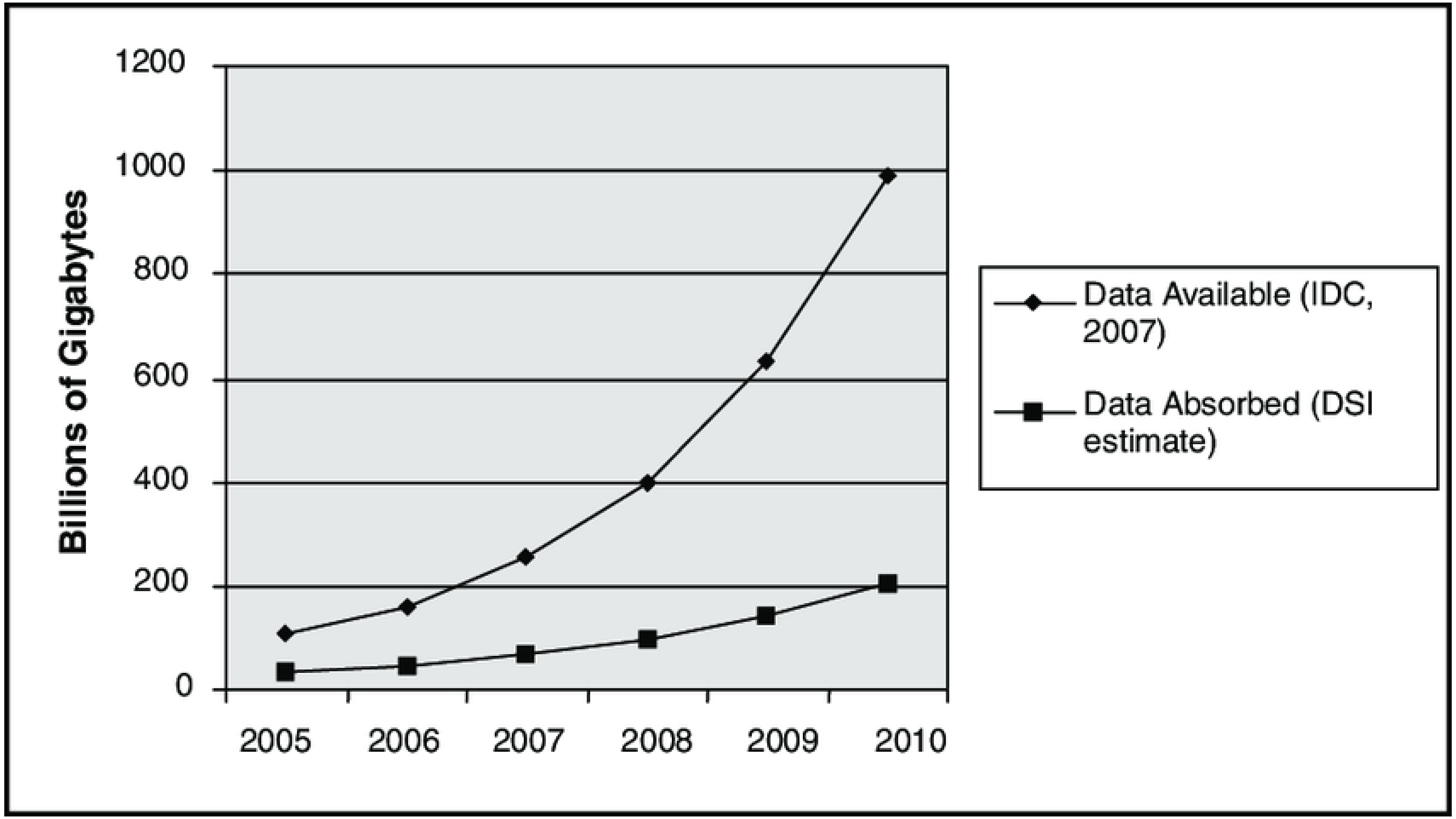
What is Bioinformatics?

- Interdisciplinary field : computer science and biology
- General Definition: A computational approach ,Solves the biological problem.

- Bioinformatics: “The collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied in molecular genetics and genomics.” (Dictionary.com)
- Bioinformatic = computational biology
- Bioinformatics, A logical and technical means by which not only solve the Biological problems but also can predicts the new aspects.
- Usually confirm whit experiment
-

Why bioinformatic is important

30% science
30% $$$$$
40% private data



The Gap is Widening…

HISTORY AND SCOPE OF BIOINFORMATICS

- By 1938, the United States Navy had developed an electromechanical analog computer small enough to use aboard a submarine.

- 1859 – The “On the Origin of Species”, published by Charles Darwin that introduced theory of genetic evolution – allows adaptation over time to produce organisms best suited to the environment.

- 1953 - Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins.

- 1955 - The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger.

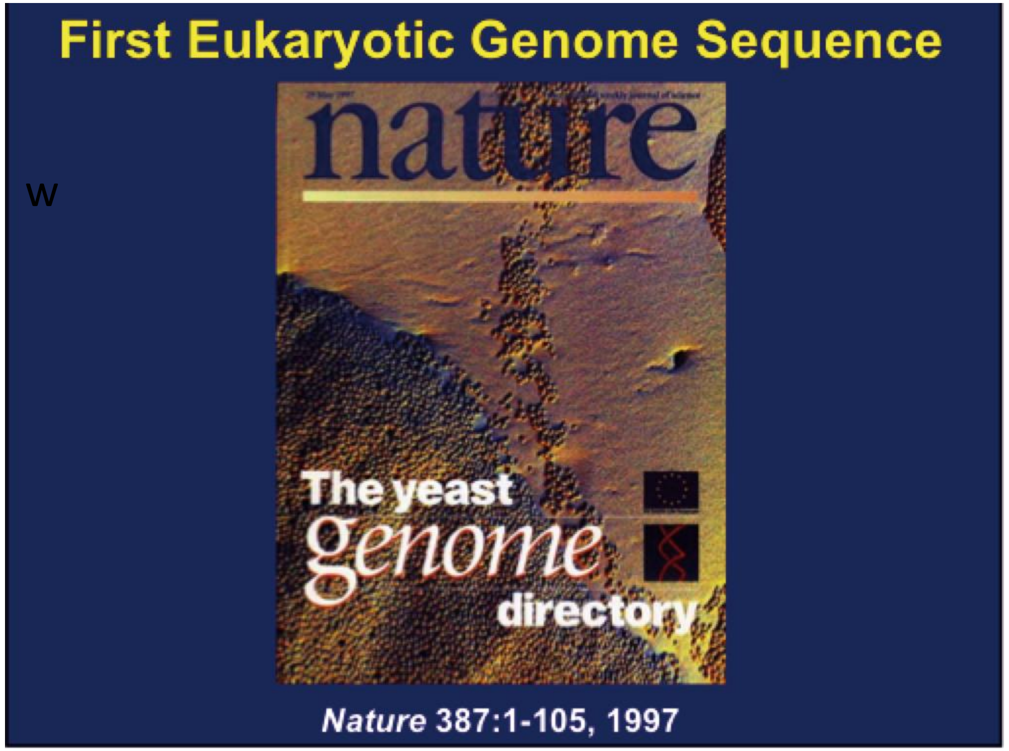
HISTORY AND SCOPE OF BIOINFORMATICS

- 1981- first modern computers
- 1990 - Human Genome Project launched
- 2001 - The human genome (3,000 Mbp) is published.
- 2010 :Completion of the 2010 Project: to understand the function of all genes within their cellular, organism and evolutionary context of Arabidopsis thaliana.

???? 2050: To complete of the first computational model of a complete cell, or maybe even already of a complete organism.

1st Eukaryotic Genome Sequence: S. cerevisiae

1st Animal Genome Sequence: C. elegans



First Eukaryotic Genome Sequence

w

Nature 387:1-105, 1997

First Animal Genome Sequence

Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

Science 282:1012-2018, 1998

February, 2001 Publications

October, 2004 Publication

articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

Nature 431:931-945, 2004

Interpreting the Human Genome Sequence

- Junk DNA
- Snps and haplotype (hapmap pr)
- Comparative genomics

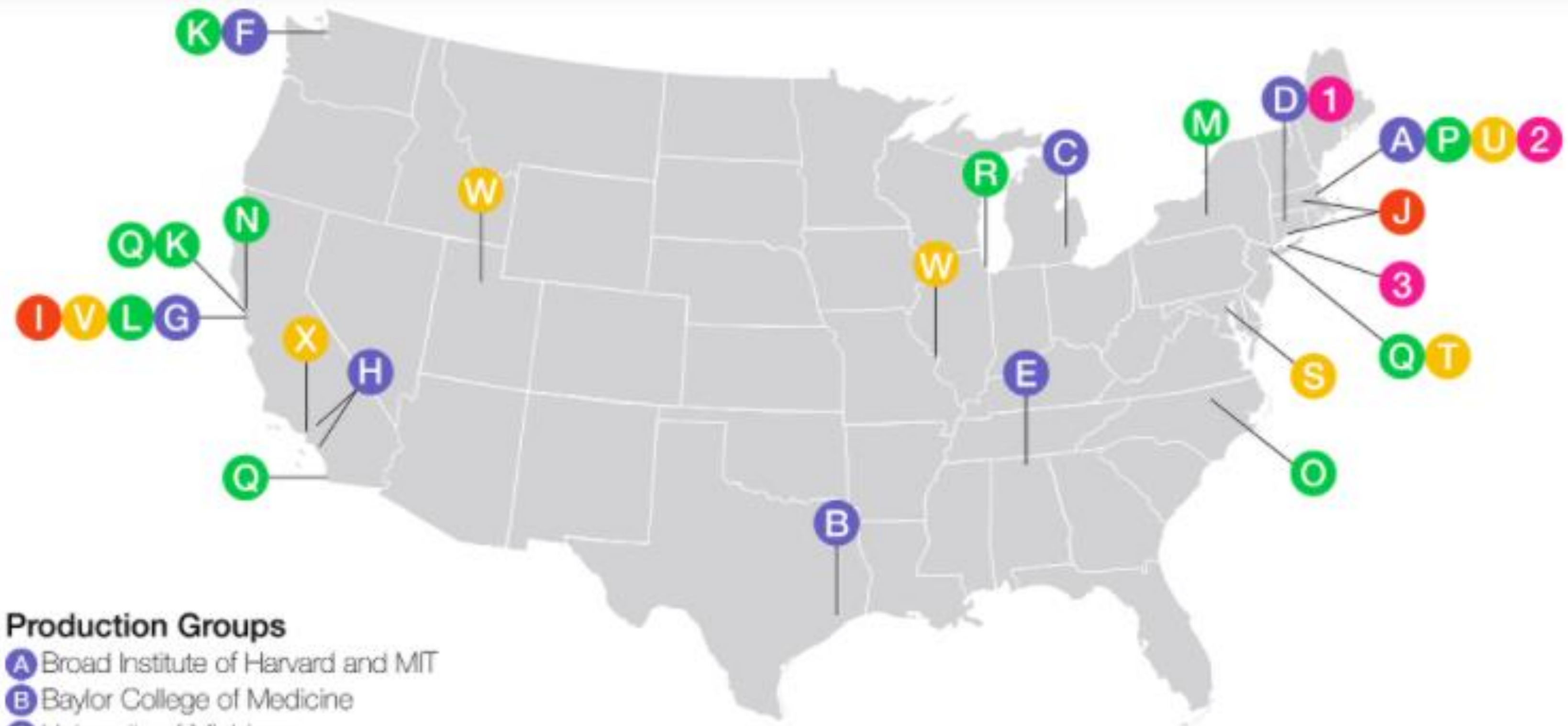
ENCODE Project

- ENCODE: Encyclopedia of DNA Elements

- The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements

- Initial pilot projrct:1% of human genom

- ENCODE investigators employ a variety of assays and methods to identify functional elements.



Production Groups

A Broad Institute of Harvard and MIT
B Baylor College of Medicine
C University of Michigan
D The Jackson Laboratory
E HudsonAlpha Institute for Biotechnology, University of Alabama in Huntsville
F Altius Institute for Biomedical Sciences
G Stanford University
H California Institute of Technology, University of California, Irvine

Data Coordination Center

I Stanford University

Data Analysis Center

J University of Massachusetts Medical School; Yale University

Characterization Centers

K University of California, San Francisco; University of Washington
L Stanford University
M Cornell University
N Lawrence Berkeley National Laboratory
O Duke University
P Broad Institute of Harvard and MIT
Q University of California, San Francisco; University of California, San Diego; Ludwig Institute for Cancer Research
R University of Chicago

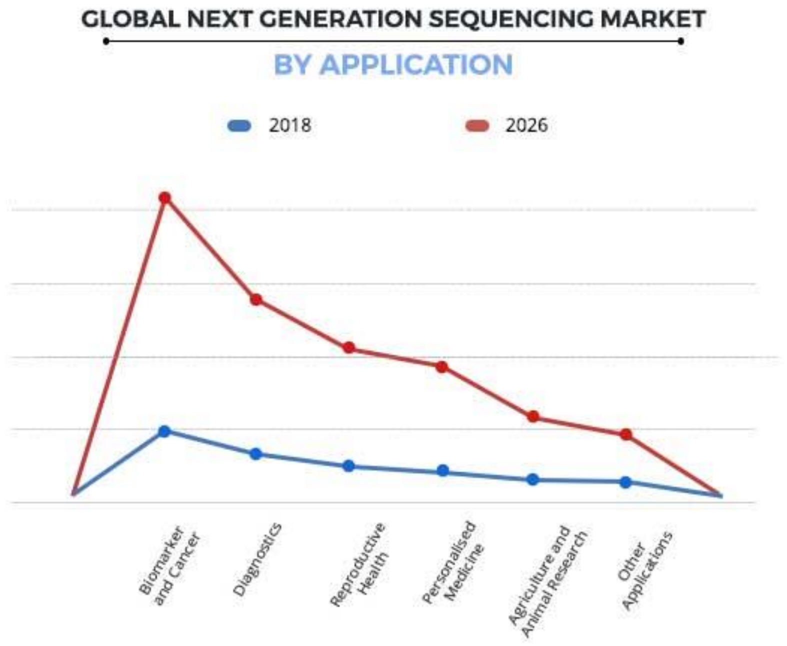
Computational Analysis Groups

S Johns Hopkins University
T Memorial Sloan Kettering Cancer Center
U Harvard University; Brigham and Women's Hospital
V Stanford University
W Washington University; University of Utah
X University of California, Los Angeles

Affiliated Groups

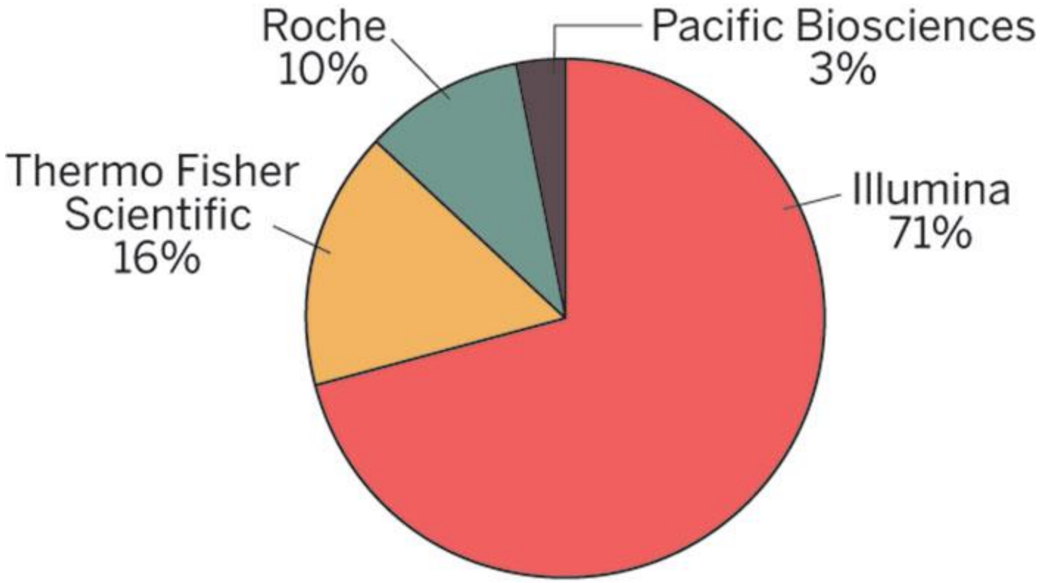
1 University of Connecticut Health Center
2 Dana-Farber Cancer Institute
3 Cold Spring Harbor Laboratory

Market of NGS sequencers

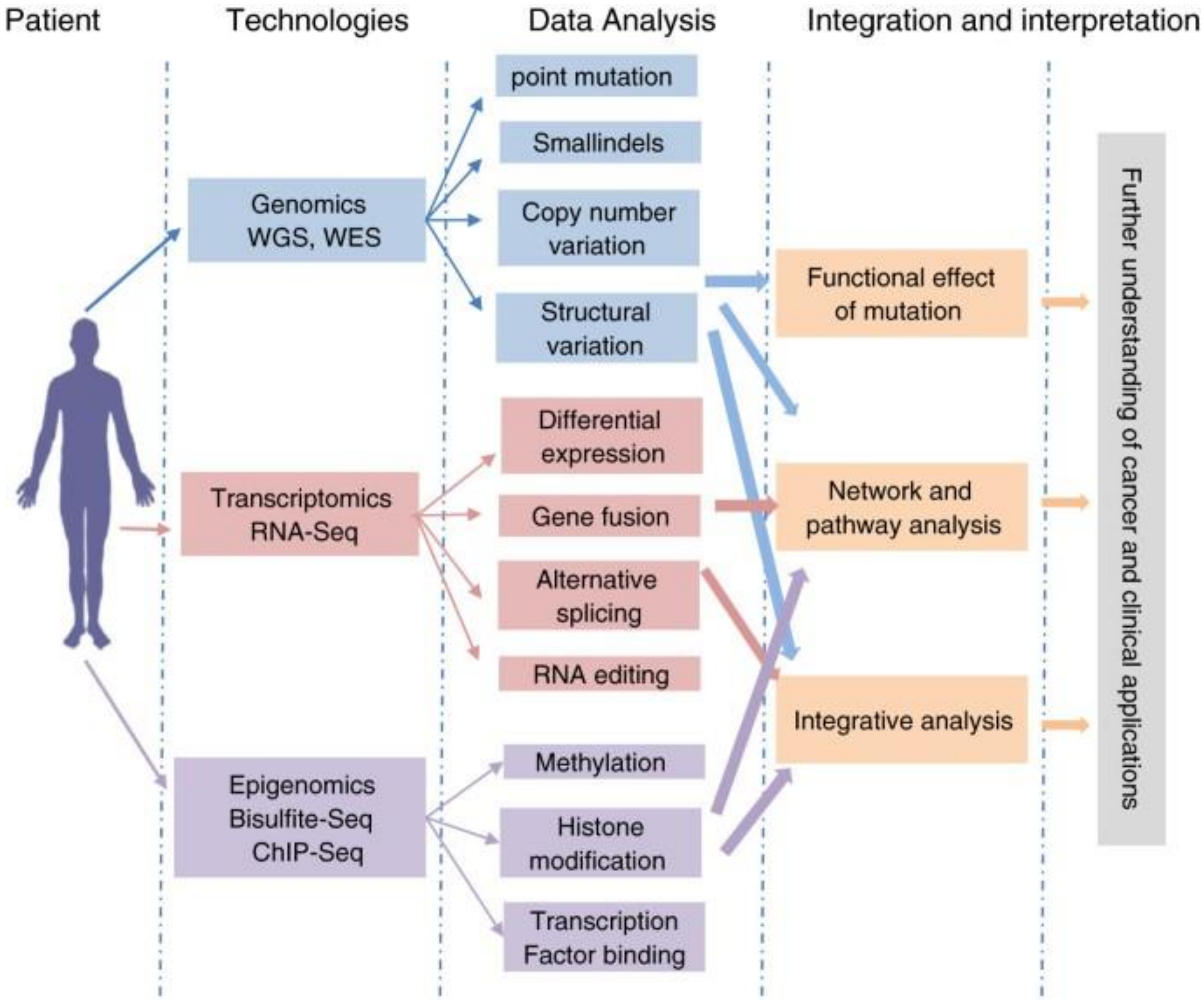

GLOBAL NEXT GENERATION SEQUENCING MARKET

BY APPLICATION

PERSONALISED MEDICINE is projected as one of the most lucrative segments.

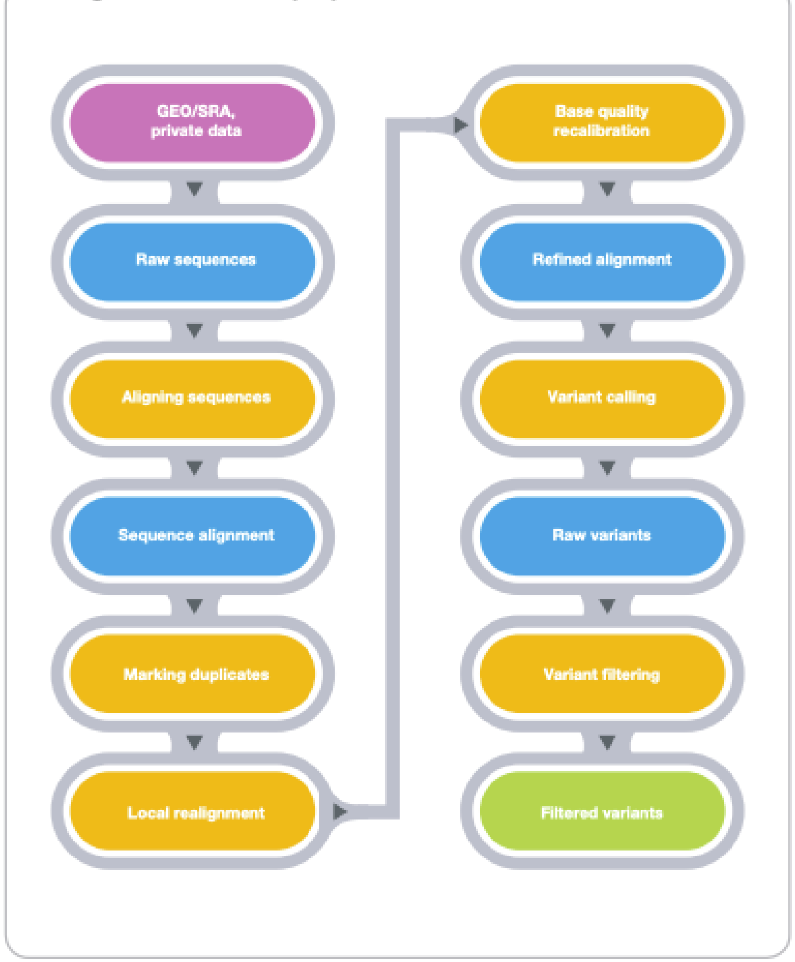


World market in 2013 = $1.3 billion


Patient
Technologies
Data Analysis
Integration and interpretation
Genomics
WGS, WES
point mutation
Smallindels
Copy number
variation
Structural
variation
Transcriptomics
RNA-Seq
Differential
expression
Gene fusion
Alternative
splicing
RNA editing
Epigenomics
Bisulfite-Seq
ChIP-Seq
Methylation
Histone
modification
Transcription
Factor binding
Functional effect
of mutation
Network and
pathway analysis
Integrative analysis
Further understanding of cancer and clinical applications

NGS DATA Processing (HGS)

Figure 1: DNA-Seq Pipeline Workflow



GEO/SRA, private data

Raw sequences

Aligning sequences

Sequence alignment

Marking duplicates

Local realignment

Base quality recalibration

Refined alignment

Variant calling

Raw variants

Variant filtering

Filtered variants

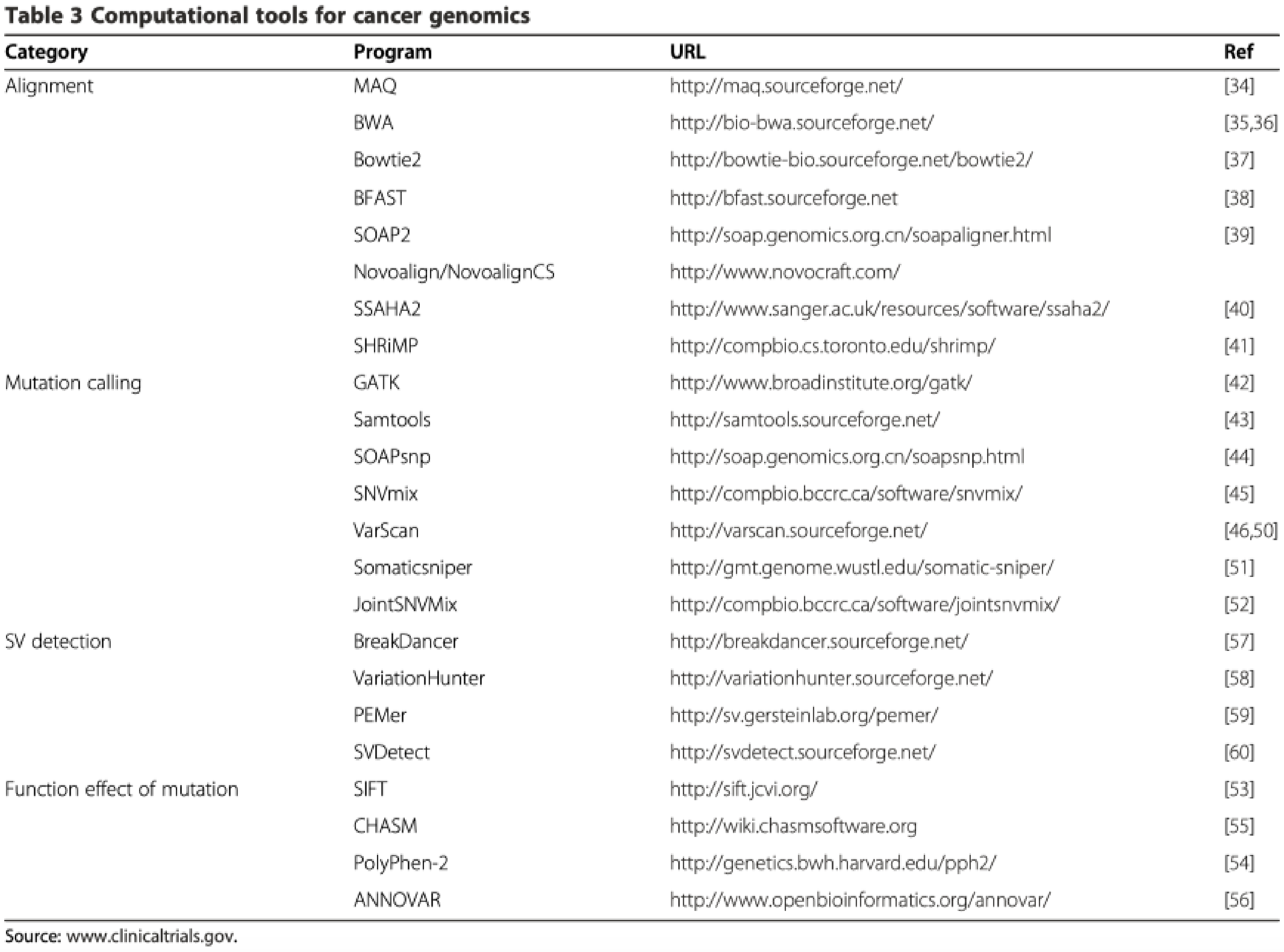


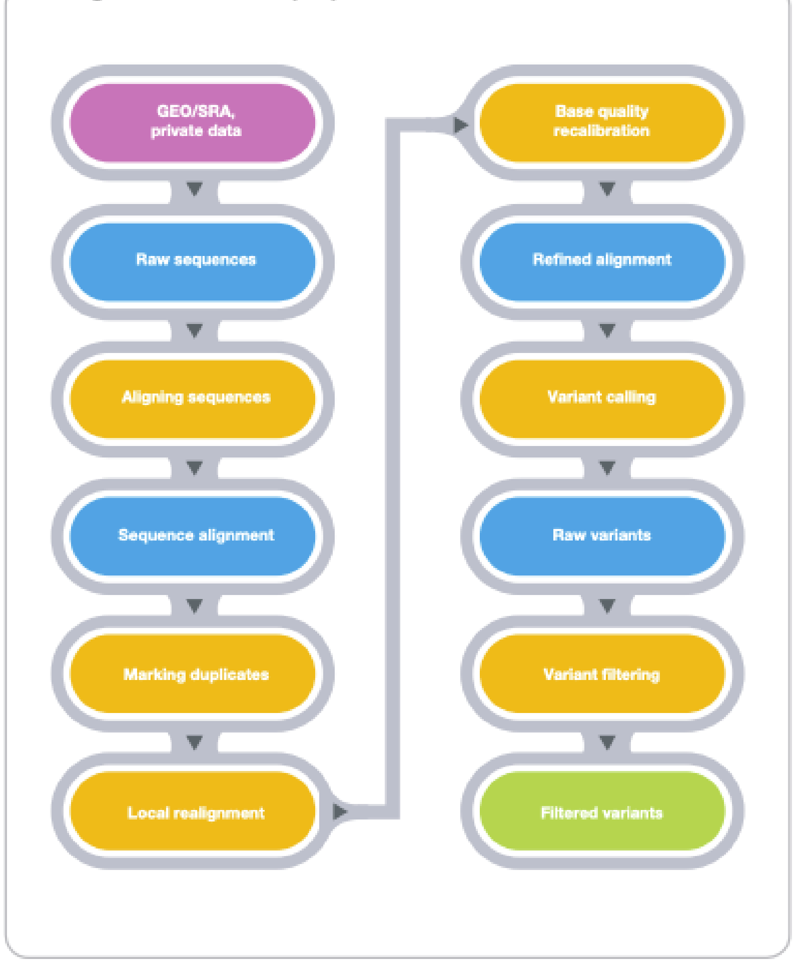
Table 3 Computational tools for cancer genomics

Category	Program	URL	Ref
Alignment	MAQ	http://maq.sourceforge.net/	[34]
	BWA	http://bio-bwa.sourceforge.net/	[35,36]
	Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/	[37]
	BFAST	http://bfast.sourceforge.net	[38]
	SOAP2	http://soap.genomics.org.cn/soapaligner.html	[39]
	Novoalign/NovoalignCS	http://www.novocraft.com/	
	SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2/	[40]
	SHRiMP	http://compbio.cs.toronto.edu/shrimp/	[41]
Mutation calling	GATK	http://www.broadinstitute.org/gatk/	[42]
	Samtools	http://samtools.sourceforge.net/	[43]
	SOAPsnp	http://soap.genomics.org.cn/soapsnp.html	[44]
	SNVmix	http://compbio.bccrc.ca/software/snvmix/	[45]
	VarScan	http://varscan.sourceforge.net/	[46,50]
	Somaticsniper	http://gmt.genome.wustl.edu/somatic-sniper/	[51]
	JointSNVMix	http://compbio.bccrc.ca/software/jointsnvmix/	[52]
SV detection	BreakDancer	http://breakdancer.sourceforge.net/	[57]
	VariationHunter	http://variationhunter.sourceforge.net/	[58]
	PEMer	http://sv.gersteinlab.org/pemer/	[59]
	SVDetect	http://svdetect.sourceforge.net/	[60]
Function effect of mutation	SIFT	http://sift.jcvi.org/	[53]
	CHASM	http://wiki.chasmsoftware.org	[55]
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/	[54]
	ANNOVAR	http://www.openbioinformatics.org/annovar/	[56]

Source: www.clinicaltrials.gov.

NGS DATA Processing (RNA seq)

Figure 1: DNA-Seq Pipeline Workflow



GEO/SRA, private data

Raw sequences

Aligning sequences

Sequence alignment

Marking duplicates

Local realignment

Base quality recalibration

Refined alignment

Variant calling

Raw variants

Variant filtering

Filtered variants

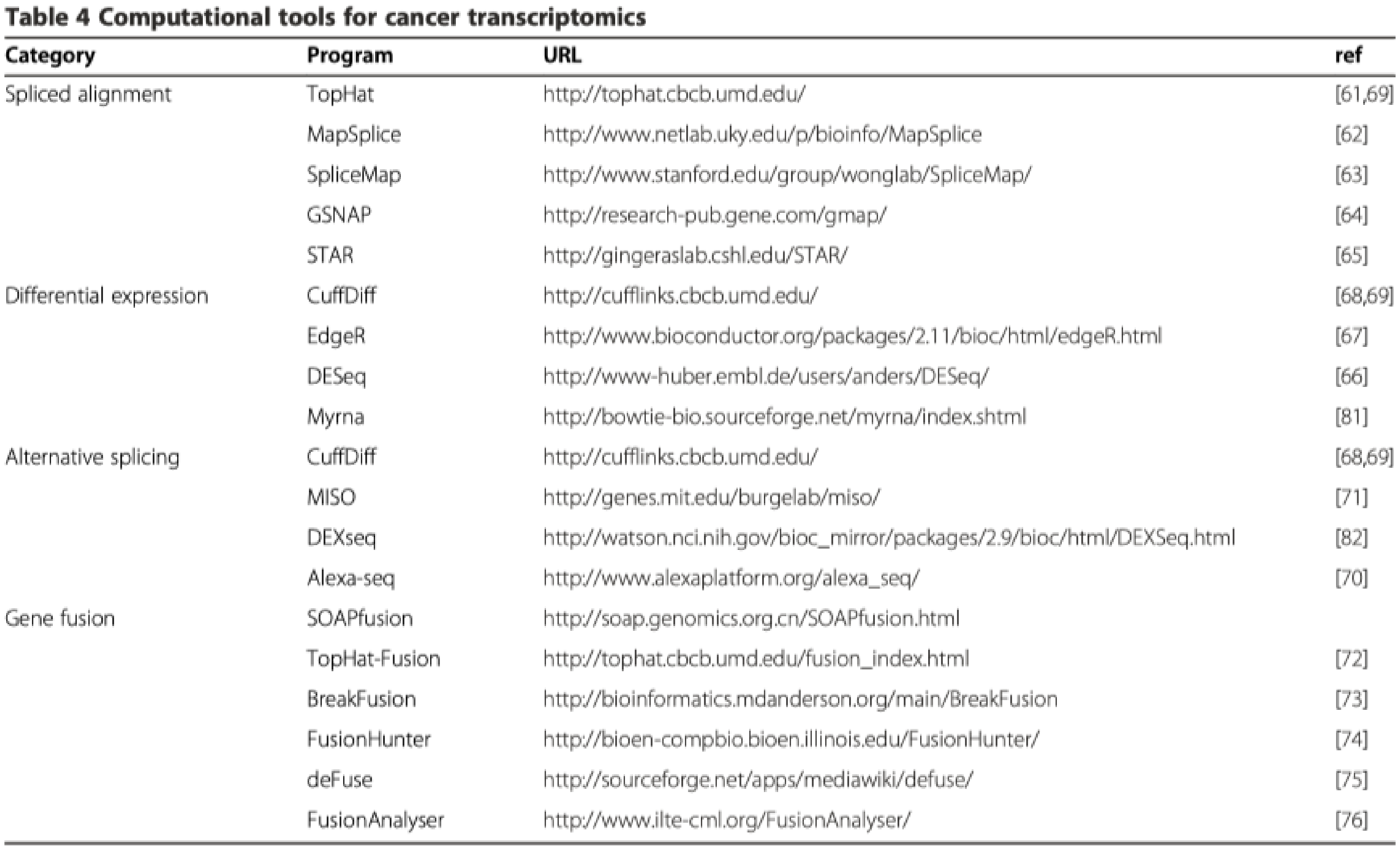


Table 4 Computational tools for cancer transcriptomics

Category	Program	URL	ref
Spliced alignment	TopHat	http://tophat.cbcb.umd.edu/	[61,69]
	MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSplice	[62]
	SpliceMap	http://www.stanford.edu/group/wonglab/SpliceMap/	[63]
	GSNAP	http://research-pub.gene.com/gmap/	[64]
	STAR	http://gingeraslab.cshl.edu/STAR/	[65]
Differential expression	CuffDiff	http://cufflinks.cbcb.umd.edu/	[68,69]
	EdgeR	http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html	[67]
	DESeq	http://www-huber.embl.de/users/anders/DESeq/	[66]
	Myrna	http://bowtie-bio.sourceforge.net/myrna/index.shtml	[81]
Alternative splicing	CuffDiff	http://cufflinks.cbcb.umd.edu/	[68,69]
	MISO	http://genes.mit.edu/burgelab/miso/	[71]
	DEXseq	http://watson.nci.nih.gov/bioc_mirror/packages/2.9/bioc/html/DEXSeq.html	[82]
	Alexa-seq	http://www.alexaplatform.org/alexa_seq/	[70]
Gene fusion	SOAPfusion	http://soap.genomics.org.cn/SOAPfusion.html	
	TopHat-Fusion	http://tophat.cbcb.umd.edu/fusion_index.html	[72]
	BreakFusion	http://bioinformatics.mdanderson.org/main/BreakFusion	[73]
	FusionHunter	http://bioen-compbio.bioen.illinois.edu/FusionHunter/	[74]
	deFuse	http://sourceforge.net/apps/mediawiki/defuse/	[75]
	FusionAnalyser	http://www.ilte-cml.org/FusionAnalyser/	[76]

Cancer genomics databases

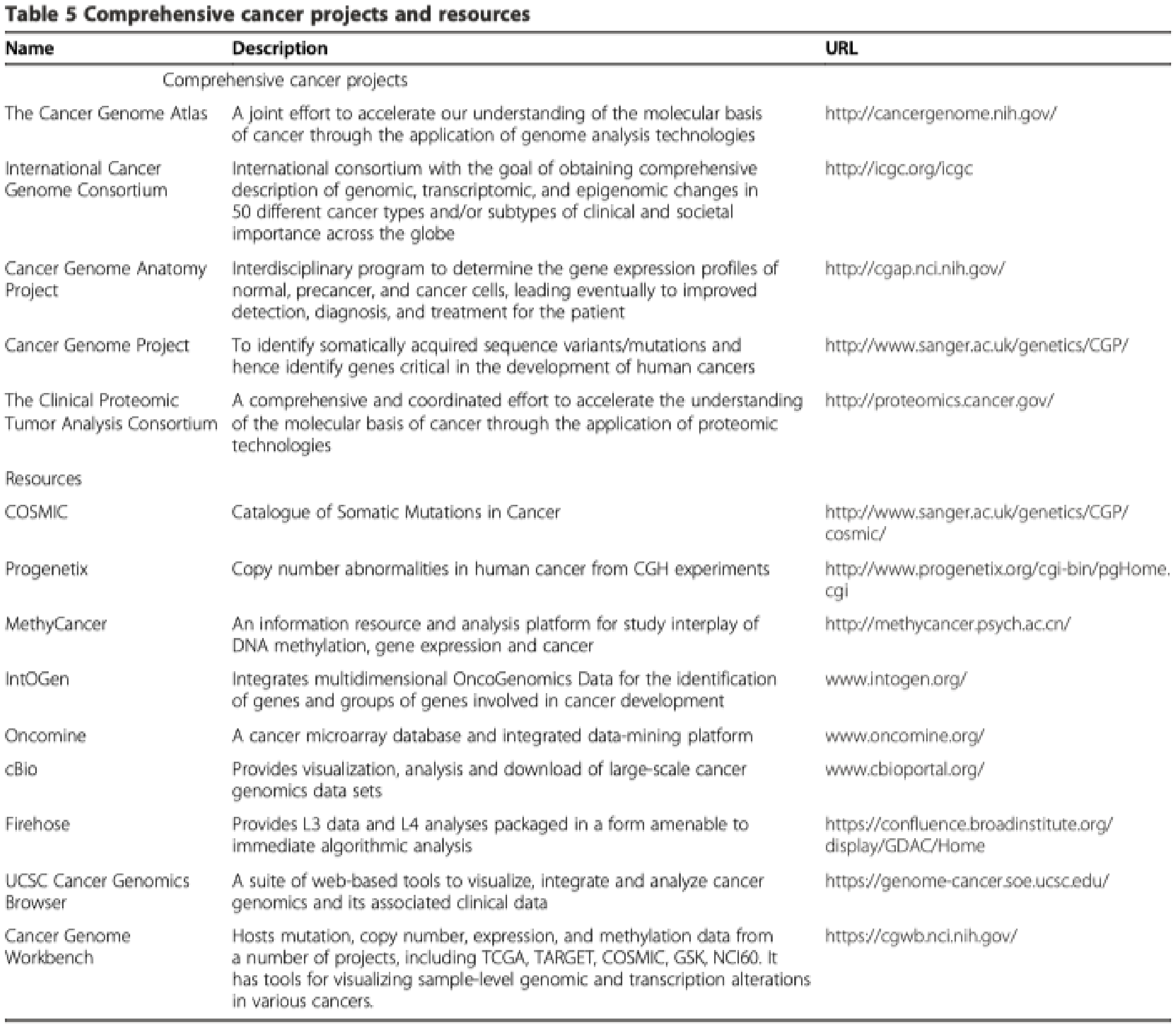
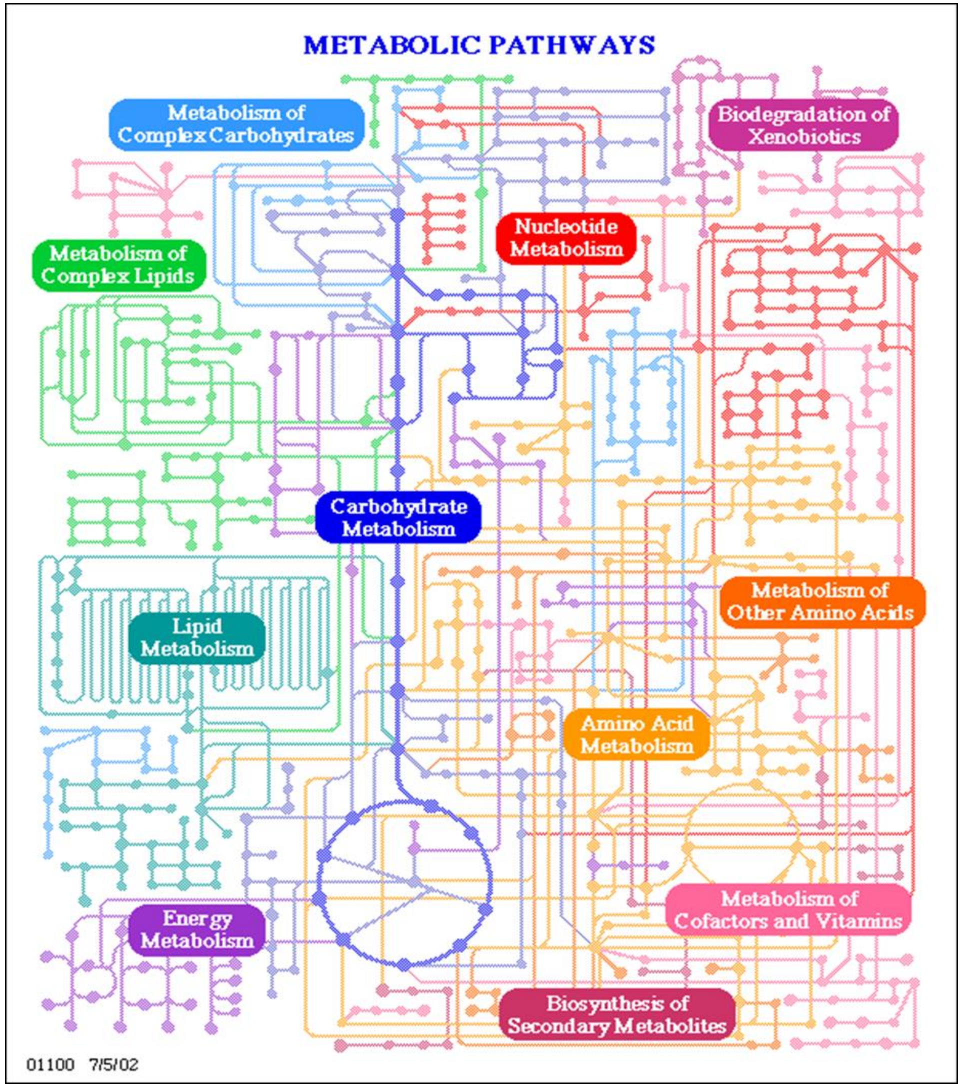


Table 5 Comprehensive cancer projects and resources

Name	Description	URL
	Comprehensive cancer projects	
The Cancer Genome Atlas	A joint effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies	http://cancergenome.nih.gov/
International Cancer Genome Consortium	International consortium with the goal of obtaining comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different cancer types and/or subtypes of clinical and societal importance across the globe	http://icgc.org/icgc
Cancer Genome Anatomy Project	Interdisciplinary program to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient	http://cgap.nci.nih.gov/
Cancer Genome Project	To identify somatically acquired sequence variants/mutations and hence identify genes critical in the development of human cancers	http://www.sanger.ac.uk/genetics/CGP/
The Clinical Proteomic Tumor Analysis Consortium	A comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of proteomic technologies	http://proteomics.cancer.gov/
Resources		
COSMIC	Catalogue of Somatic Mutations in Cancer	http://www.sanger.ac.uk/genetics/CGP/cosmic/
Progenetix	Copy number abnormalities in human cancer from CGH experiments	http://www.progenetix.org/cgi-bin/pgHome.cgi
MethyCancer	An information resource and analysis platform for study interplay of DNA methylation, gene expression and cancer	http://methycancer.psych.ac.cn/
IntOGen	Integrates multidimensional OncoGenomics Data for the identification of genes and groups of genes involved in cancer development	www.intogen.org/
Oncomine	A cancer microarray database and integrated data-mining platform	www.oncomine.org/
cBio	Provides visualization, analysis and download of large-scale cancer genomics data sets	www.cbioportal.org/
Firehose	Provides L3 data and L4 analyses packaged in a form amenable to immediate algorithmic analysis	https://confluence.broadinstitute.org/display/GDAC/Home
UCSC Cancer Genomics Browser	A suite of web-based tools to visualize, integrate and analyze cancer genomics and its associated clinical data	https://genome-cancer.soe.ucsc.edu/
Cancer Genome Workbench	Hosts mutation, copy number, expression, and methylation data from a number of projects, including TCGA, TARGET, COSMIC, GSK, NCI60. It has tools for visualizing sample-level genomic and transcription alterations in various cancers.	https://cgwb.nci.nih.gov/

SYSTEMS BIOLOGY

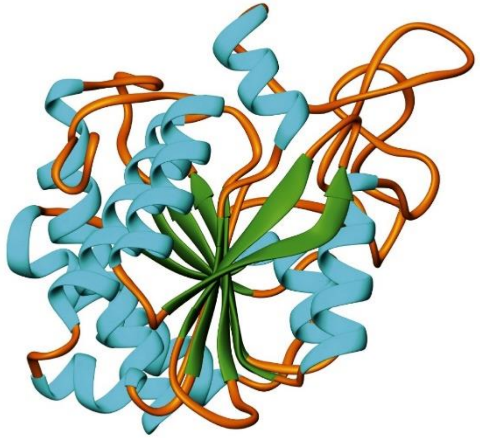
- Gene gene interaction
- Protein gene interaction
- Metabolite (drug) –biomolecule interaction
- Dynamic of gene regulation
- Triggering- gene
- Hub-genes


METABOLIC PATHWAYS
Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 7/5/02

In silico structural biology

In silico structural biology

- INPUT: (PDB FILES) from pdb database
- RNA
- DNA
- LIPIDS(membrane)
- PROTEINS



© 2008 John Wiley & Sons, Inc. All rights reserved.

In silico methods, ranging from bioinformatics analysis of primary sequences, through computer simulations of tertiary structures, to the prediction of novel structures by de novo design, wind through the platforms aimed at constructing optimal biocatalysts

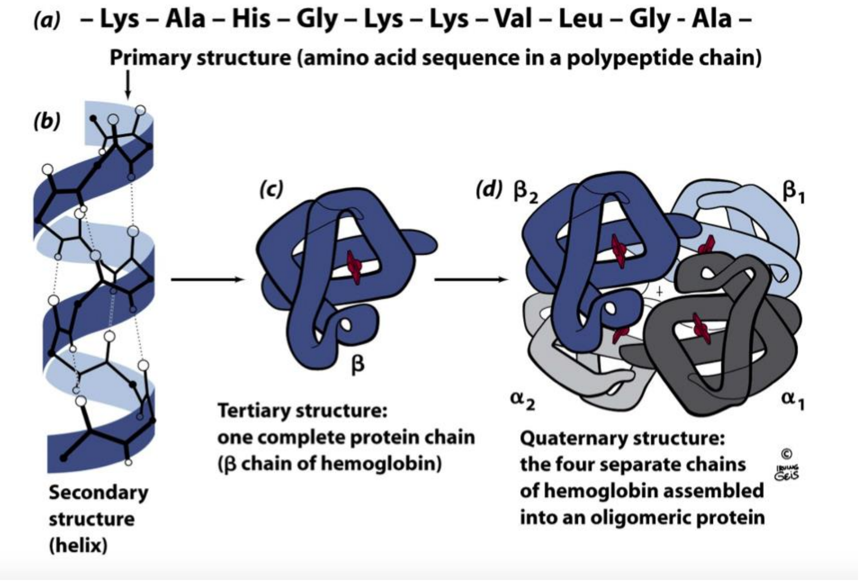
PROTEIN FOLDING

1. Primary structure: amino acid sequence

2. Secondary structure: local arrangement of peptide backbone

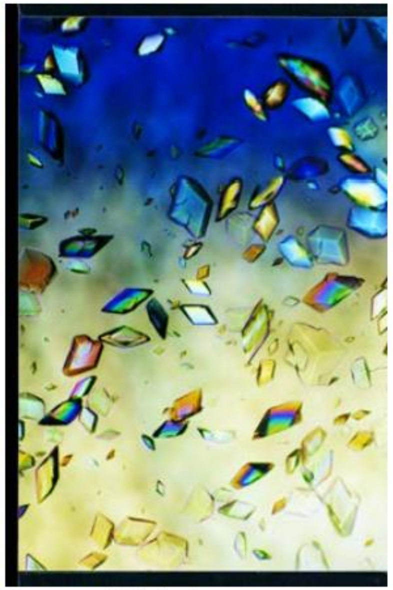
3. Tertiary structure: three dimensional arrangement of all atoms, peptide backbone and amino acid side chains

4. Quaternary structure: spatial arrangement of subunits

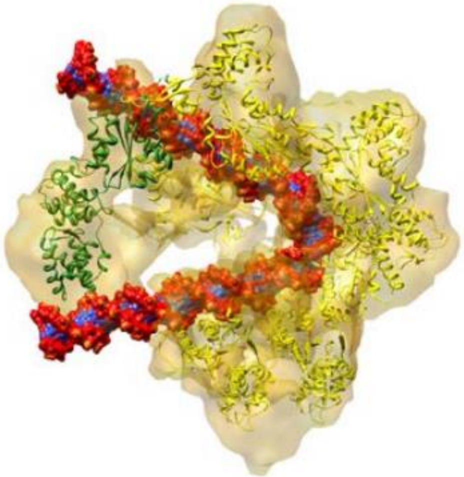

(a) – Lys – Ala – His – Gly – Lys – Lys – Val – Leu – Gly - Ala –
Primary structure (amino acid sequence in a polypeptide chain)
(b)
Secondary structure (helix)
(c)
β
Tertiary structure: one complete protein chain (β chain of hemoglobin)
(d) β_2
β_1
α_2
α_1
Quaternary structure: the four separate chains of hemoglobin assembled into an oligomeric protein
©

What is structural biology?

- Structural biology is the study of how biological molecules are built. Using a variety of imaging techniques, scientists view molecules in three dimensions to see how they are assembled, how they function, and how they interact (Structure and function relationships)



crystals of proteins:
University of California, Irvine.



Cryo-EM: Bruce Stillman,
Cold Spring Harbor Laboratory.

900 megahertz NMR Systems.

Can scientists view how proteins act?

- New technology is beginning to allow researchers to progress from creating static pictures of proteins and other molecules to making movies of their actions.
- Images provide snapshots of what these cellular elements are doing at specific points in time.

❖ Although they supply valuable information, these still pictures don’t capture how proteins and other molecules inside cells are constantly moving and changing, folding and unfolding as they interact.

How can proteins fold so fast?

- Proteins fold to the lowest-energy fold in the microsecond to second time scales. How can they find the right fold so fast?

- It is mathematically impossible for protein folding to occur by randomly trying every conformation until the lowest-energy one is found (Levinthal’s paradox)

- Search for the minimum is not random because the direction toward the native structure is thermodynamically most favorable

two main approaches to design proteins

- two main approaches are used to design the novel proteins or enzymes: rational design and directed evolution

- rational design via site directed or saturation mutagenesis and directed evolution via random mutagenesis are used as key tools in protein engineering.

- The preliminary knowledge of protein structure is not required in directed protein evolution

RATIONAL DESIGN

- PROTEIN FUSION: domain and motif shuffeling ((FRANKSHTAIN PROTOCOL)
- Novel function
- STABLE PROTEINS
- PEPTIDE DESIGN (anticancer peptides , antimicrobial peptides)
- anticancer peptides : tumor homing – penetrating
- Vaccine design : epitope mapping - humanized antibody
- Receptor based drug discovery

Big problem: Homology or and novo modelling

crystallography or NMR IS impossible

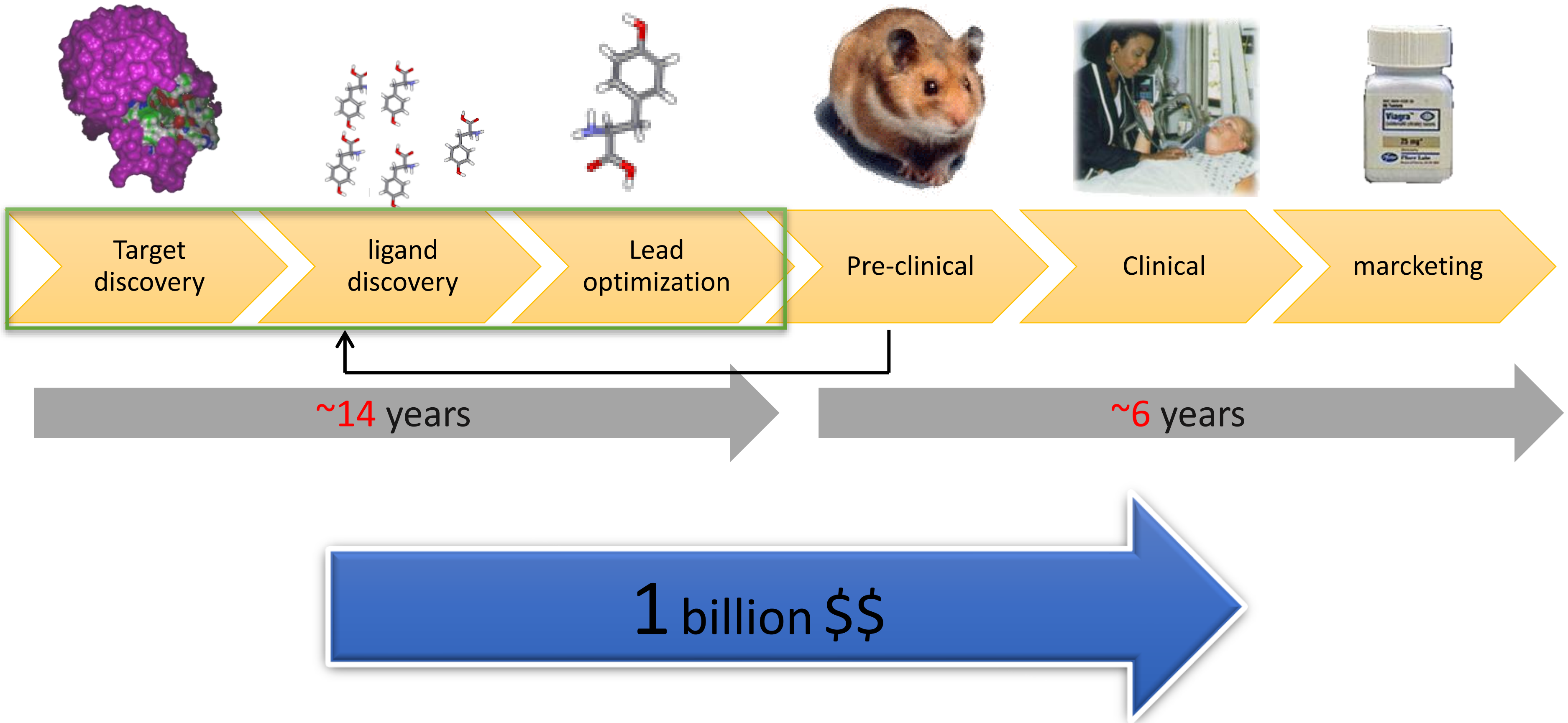
- Mutagens vs wildtype
- Humanized antibody
- Domain extraction
- Fusion and Tags

- Rifine whit MD

DRUG DISCOVERY

DRUG DISCOVERY

Receptor based or QSAR



Target discovery

ligand discovery

Lead optimization

Pre-clinical

Clinical

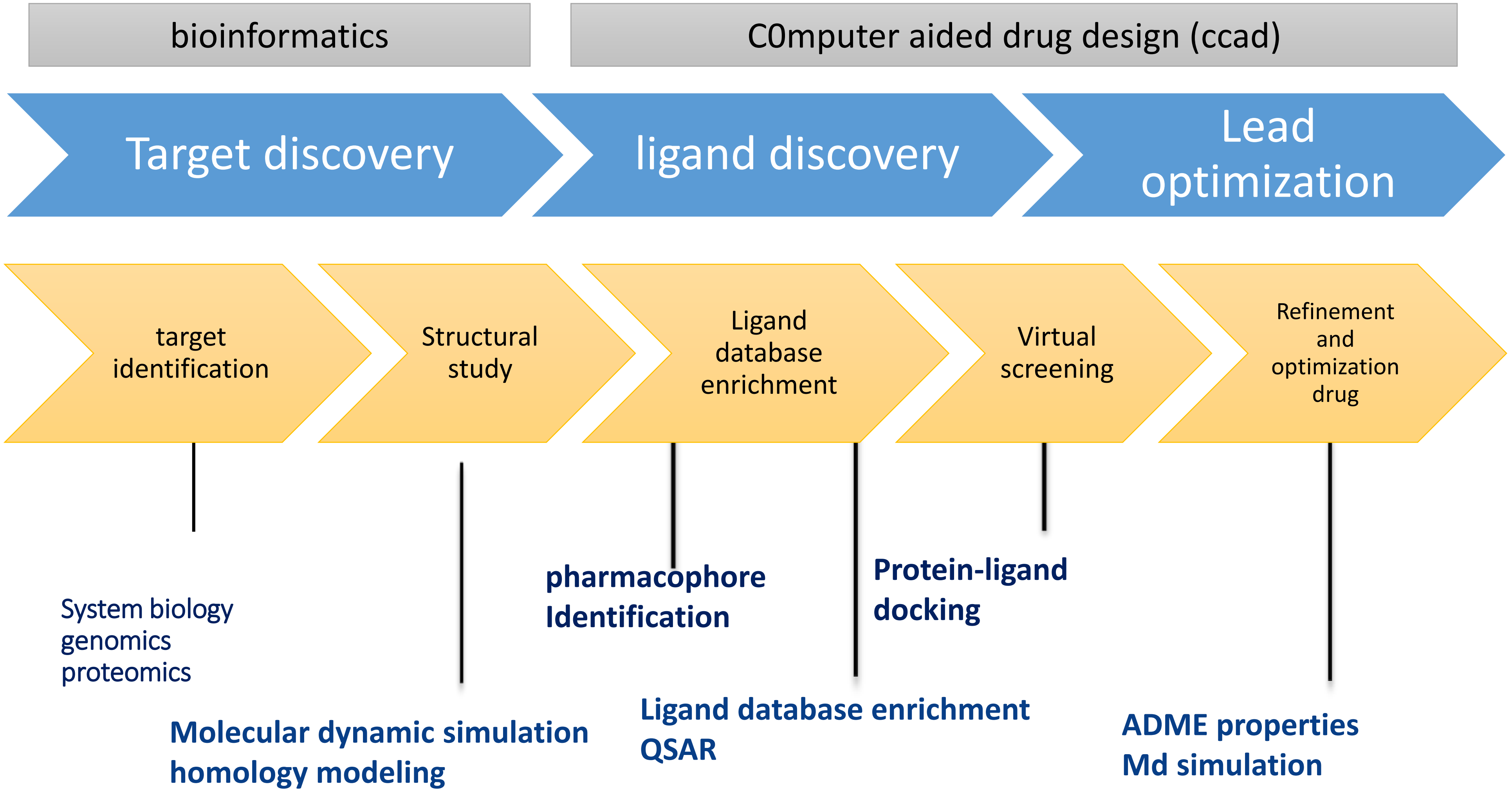
marcketing

~14 years

~6 years

1 billion $$

DRUG DISCOVERY



bioinformatics

C0mputer aided drug design (ccad)

Target discovery

ligand discovery

Lead optimization

target identification

Structural study

Ligand database enrichment

Virtual screening

Refinement and optimization drug

System biology
genomics
proteomics

Molecular dynamic simulation
homology modeling

pharmacophore
Identification

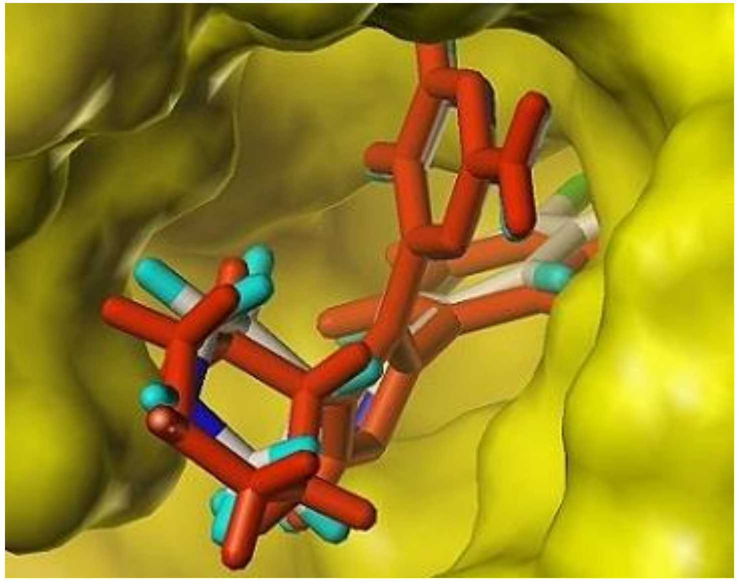
Ligand database enrichment
QSAR

Protein-ligand
docking

ADME properties
Md simulation

What is Docking?

- Molecular docking studies are used to determine the interaction of two molecules and to find the best orientation of ligand which would form a complex with overall minimum energy



Docking

- Can be applied to many systems:

- Protein – small molecule
- Protein – protein
- Protein – DNA
- Small molecule – DNA
- auto assemble (virus capsid)

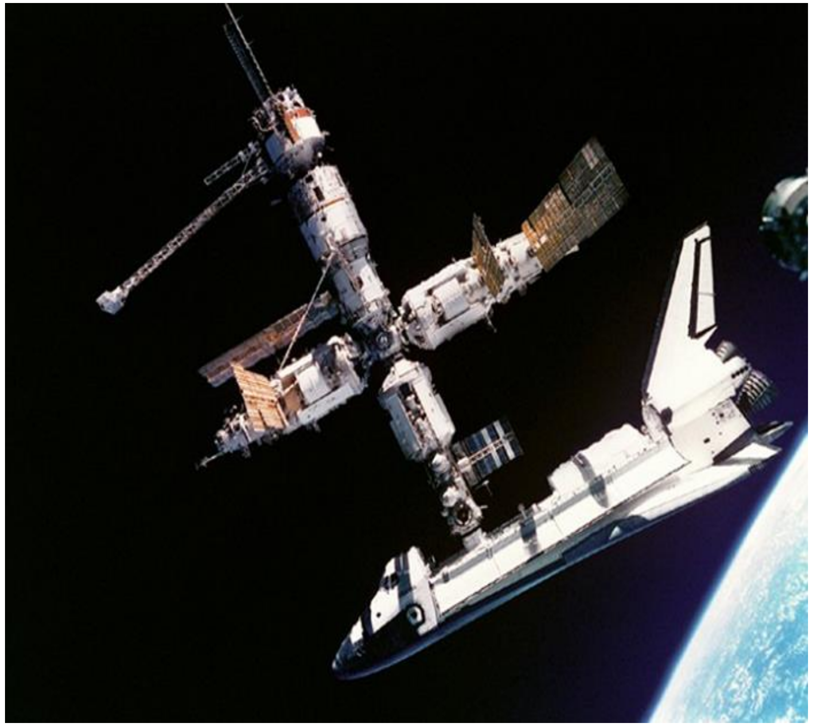
Rigid docking

❖Key & lock

- Six degree of freedom:
- Protein and ligand both treated as rigid
- 3 rotation/3 transition

Solution : Rotamer library

- Just like docking the space shuttle whit satellite



MD Simulations

- Very valuable for understanding the dynamic behavior of proteins at different time scales ,from fast internal motions to slow conformational changes or even protein folding processes

- Possible To Study the effect Of Explicit Solvent Molecules On Protein Structure and stability to obtain time-averaged Properties Of The Bimolecular system

- X-ray or NMR or models methods have been refined using MD methods

MD Simulations

- first protein MD simulation(bovine pancreatic trypsin inhibitor; 58 residues and 450atoms) was done in vacuum and for only 8.8 psec

- nowadays permit simulations of systems comprising 10000–10000000 atoms and simulation times in the order of nsec to $m$sec Simulations of more realistic systems, including explicit water molecules

- descriptors for non-standard molecules, such as ligands, might be missing

Combined Docking and MD Simulations

- Fast and inexpensive docking protocols can be combined with accurate but more costly MD techniques to predict more reliable protein–ligand complexes

- The strength of this combination lies in their complementary strengths and weaknesses

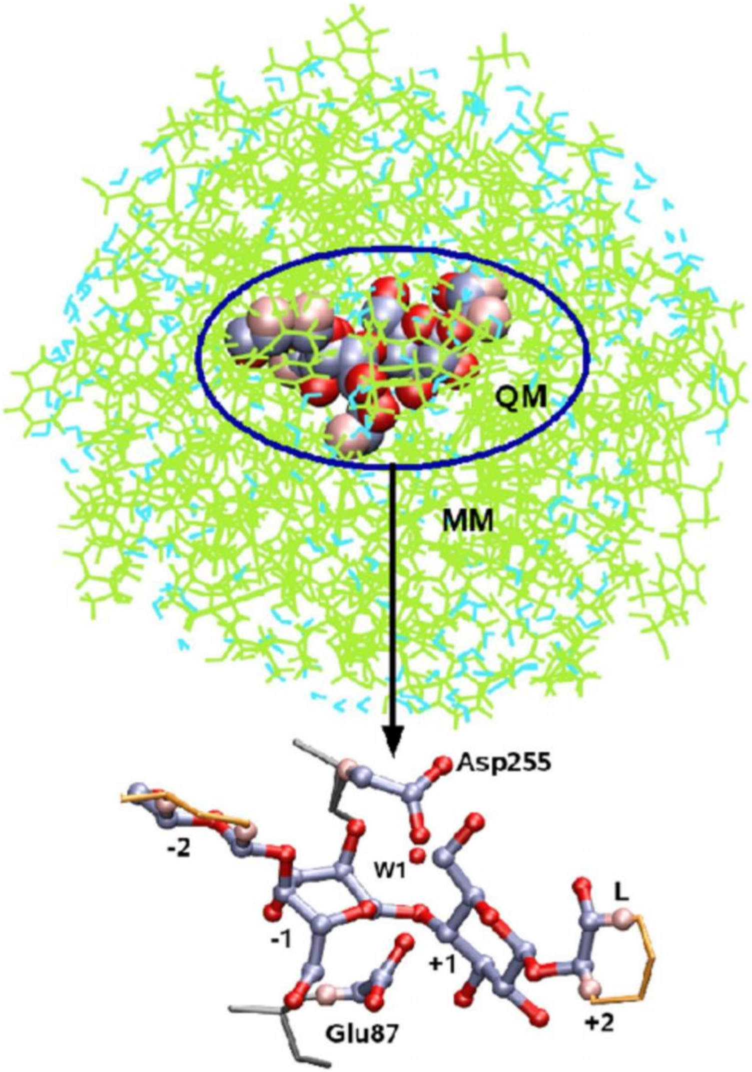
- One the one hand, docking techniques are used to explore the vast conformational space of ligands in a short time

- On the other hand ,MD simulations can treat both ligand and protein in a flexible way, allowing for an induced fit of the receptor-binding site around the newly introduced ligand.

- In addition, the effect of explicit water molecules can be studied directly, and very accurate binding free energies can be obtained.

MM-QM

- Molecular mechanic vs Quantom mechaic
- Schrodinger eq
- catalytic mechanism
- electrons



other topics

- DATA mining : statistic , programing skils, machin learning
- Systems biology
- Proteomics
- Metabolomics and chemometrics
- Lipidomics
- Chemoinformatics
- And……………………….

Brief Answers to the Big Questions

What kind of bioinformatics analysis need supercomputers and whit kind just need laptops

Best laptop for bioinformatics?

CPU and GPU base computing

Cloud computing

Rent a server

Brief Answers to the Big Questions

- Linux, mac or windows?

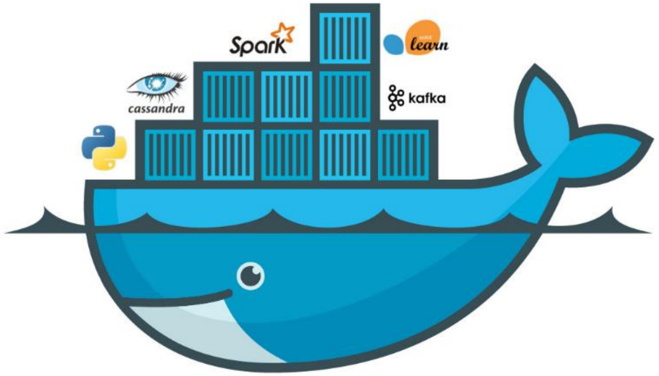
- CLI vs GUI

Brief Answers to the Big Questions

- which programing language should i learn first?
- Perl
- Python
- R
- BASH

Some learning boosters

- Git and git-hub
- Anaconda and byoconda
- Bioconductor
- Maling list and
- DOCKER



Learning materials

- Books
- Websites (tutorials)
- Git-hub

Online courses

Coursera

edX

- YOUTUBE

ByOMICs

- Linux course
- R course
- Python course
- Protein design
- Drug discovery
- NGS
- Data mining
- Metabolmics and ……..

Thanks…