High-Performance Computing: the example of Merck Serono

Friedrich Rippmann, Bio- and Chemoinformatics
rippmann@merck.de

Introducing Merck Serono

- Merck is the oldest pharmaceutical and chemical company in the world
- Establishment of the new Merck Serono division in early 2007 following the acquisition of Serono S.A. by Merck; division now headquartered in Genevea
- Focus on innovative, prescription drugs of chemical and biological origin
- Major research sites in Darmstadt, Geneva and Boston
- R&D Investment: € 1 billion (2008)
- R&D FTEs: 2300
- >100 clinical trials ongoing
- over 150 active external collaborations (former Research)

The Pützer Tower and Pyramid at Darmstadt Merck headquarters

Merck Serono headquarters in Geneva

HPC at Merck, Merck Serono

Awareness phase	Test & establ. phase	Production phase
1991	1996	2001
CONVAX (in-house)	Specialized HW (external) e.g. Parsytec [PC Cluster (internal)]	IBM & SGI clusters & multiprocessor machines
calculation speed	calculation speed	calculation & storage sp.
limited in-house activity	academic collaborations: MAXHOM, PHASE, TTN, TARGID…	all in-house
molecular dynamics	sequence annotation, docking, structure prediction	sequence annotation, docking, genome-wide association studies, large-scale clustering
low project impact	low project impact	good project impact

HPC – external or internal?

	External	Internal
pro	- up-to-date hardware - on demand (potentially lower cost) - (almost) unlimited capacity	- safe - flexible (you install & run whatever software whenever you need it)
con	- Software installation - Data transmission - Security: what happens if service leaks (academia), or commercial service providers goes bankrupt (and your data are sold together with the machines…)?	- high investment, maintenance - know-how intensive (load balancing systems, memory allocation etc.)

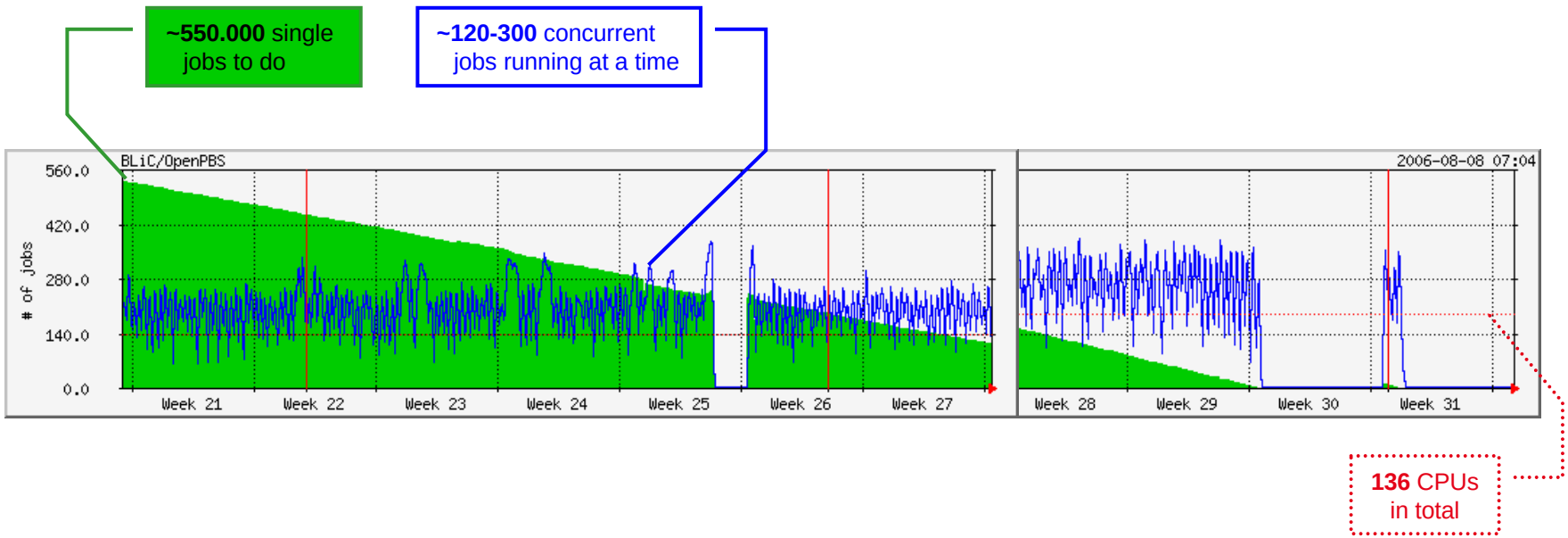
retained option (for now)

Application „sequence annotation“

- Various tasks (early application: GENEQUIZ (1996); today e.g. re-annotation of Affymetrics probe sets)
- Typical run times: days to weeks
- Software used: e.g. BLAST, in-house SW
- Issues: data volumes, storage speed
- Project impact: medium (must be done, but no immediate discovery)

Example re-annotation of GeneChip™ probes

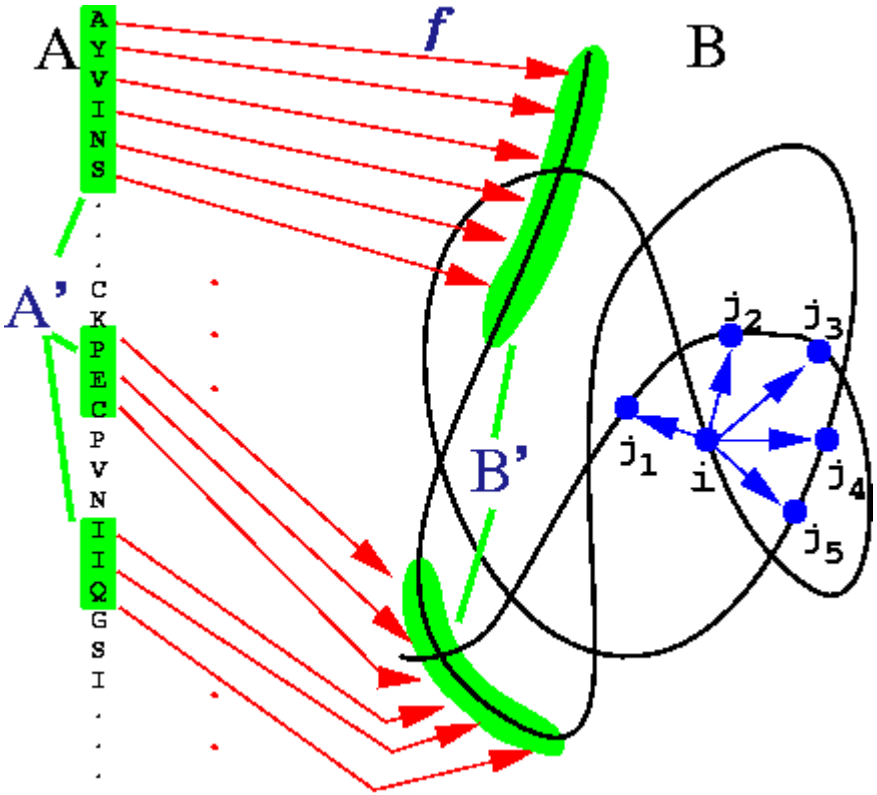
Run times of several weeks for a single annotation job

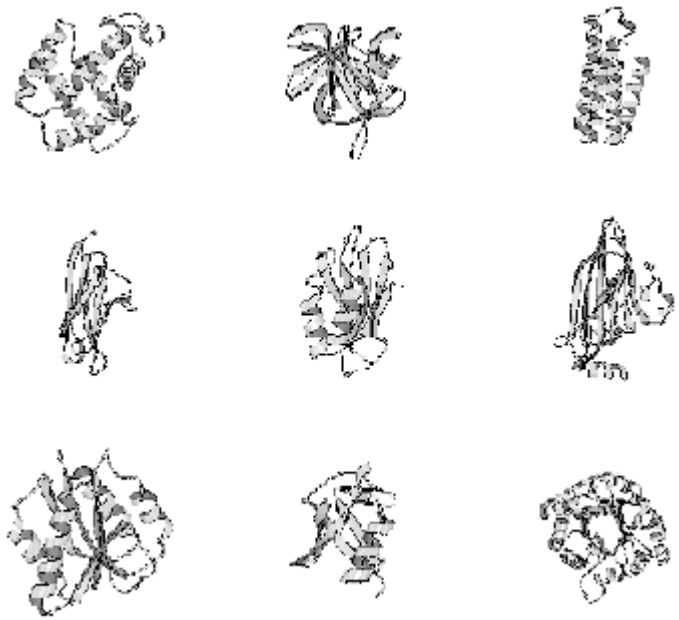


Application „threading“

- A given target sequence is „threaded“ on all known protein folds, with gaps of any length at any position
- Typical run time: several days (evaluation 1-2 weeks)
- Software used: THREADER, 123D and other
- Issues: results often not conclusive
- Project impact: very low

Fold prediction via ´threading´





A given sequence...

...is threaded in all positions...

...onto all known folds

BMBF Project Targid

Application „*ab initio* prediction“

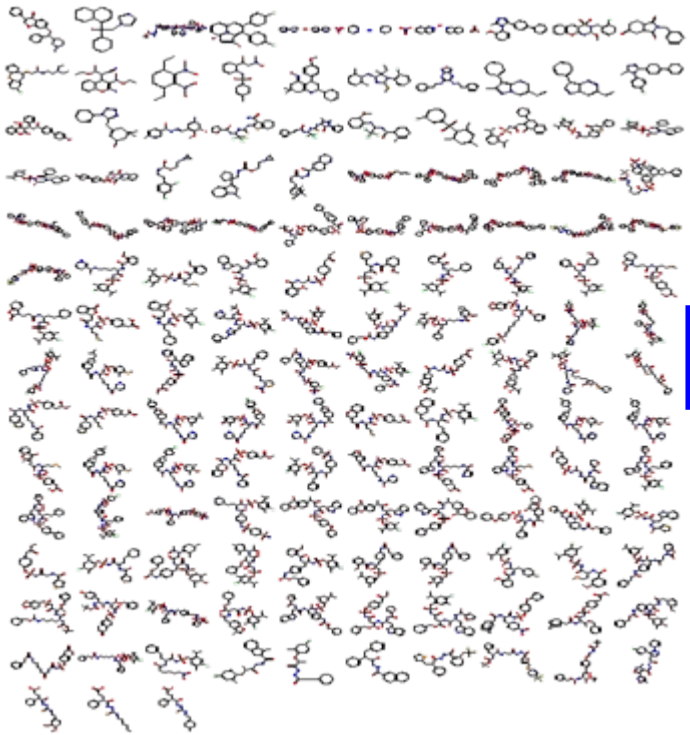
- A given target sequence is „folded“ into ist likely 3D structure
- Typical run time: several hours
- Software used: DRAGON (Aszodi & Taylor)
- Issues: works only on small proteins
- Project impact: very low

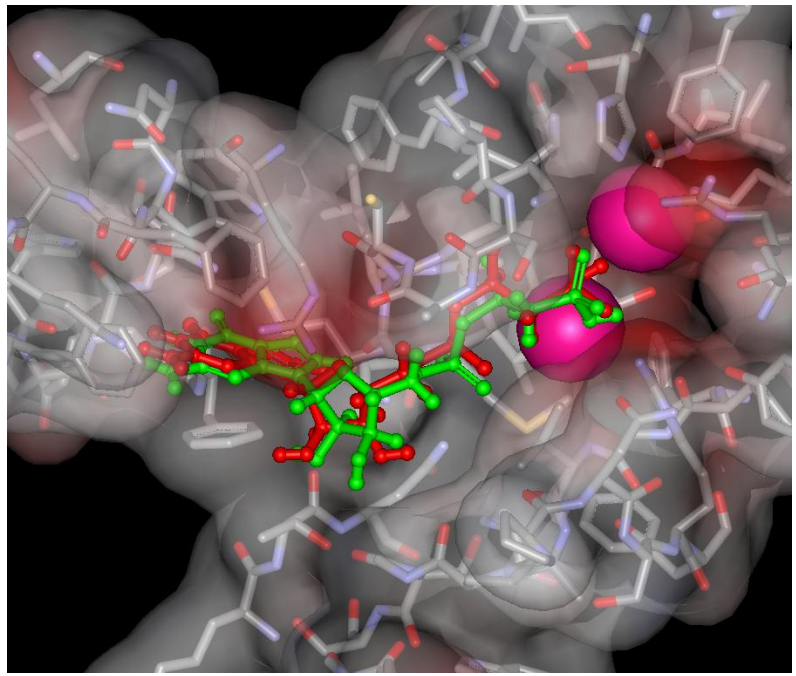
Application „docking“

- Large numbers of chemical structures are docked into a predefined cavity of the protein target
- Typical run time: 1-2 days (evaluation 1-2 weeks)
- Software used: FLEXX, GOLD
- Issues: large numbers of files generated cause problems to file systems and their operation (e.g. listing, deletion of a few result sets takes several days)
- Project impact: very high (has regularly generated new guide structures for chemists)

Virtual screening by ligand docking

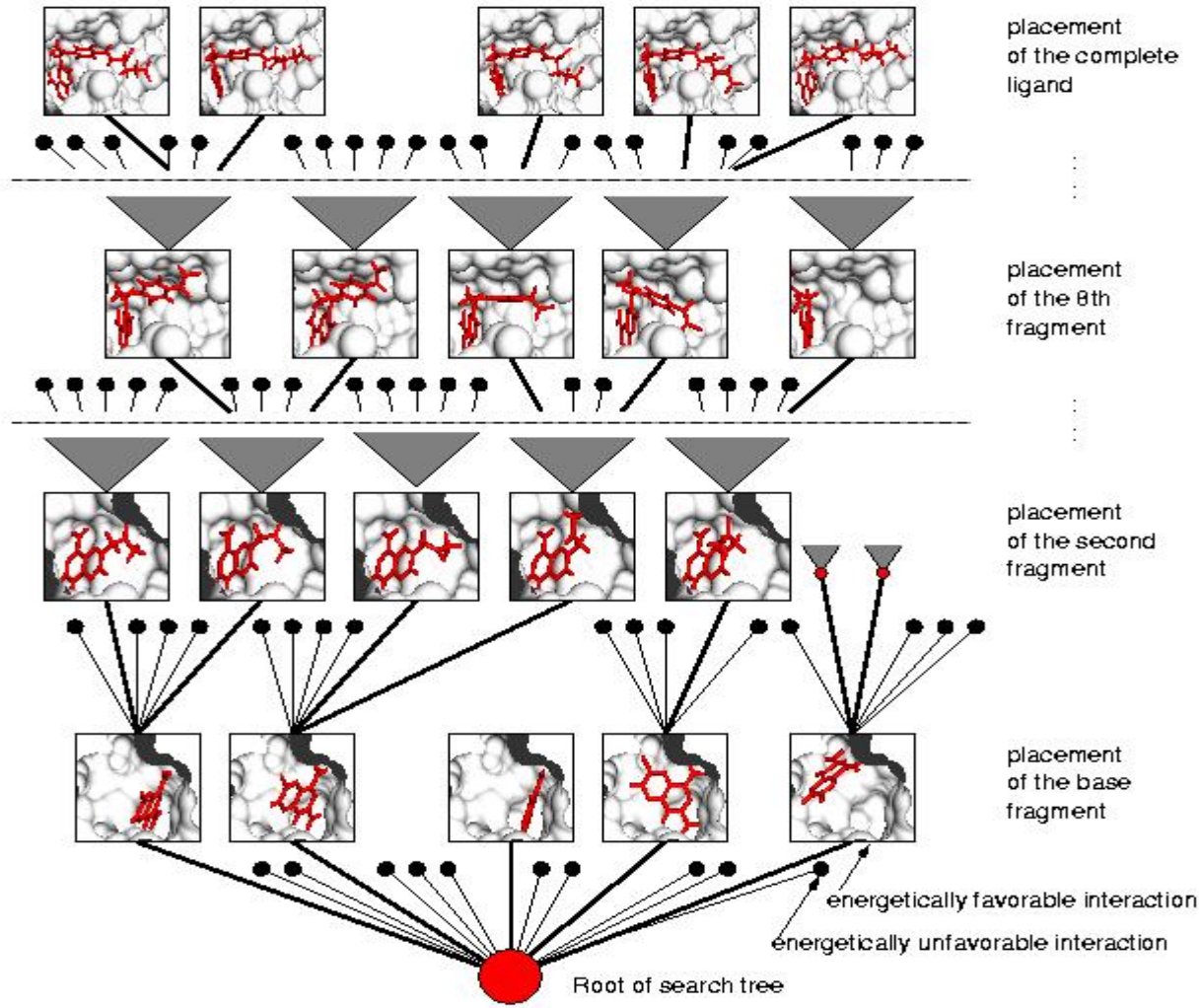
Find *in silico* suitable molecules for my target





take large numbers of structures **calculate** **find those that fit**

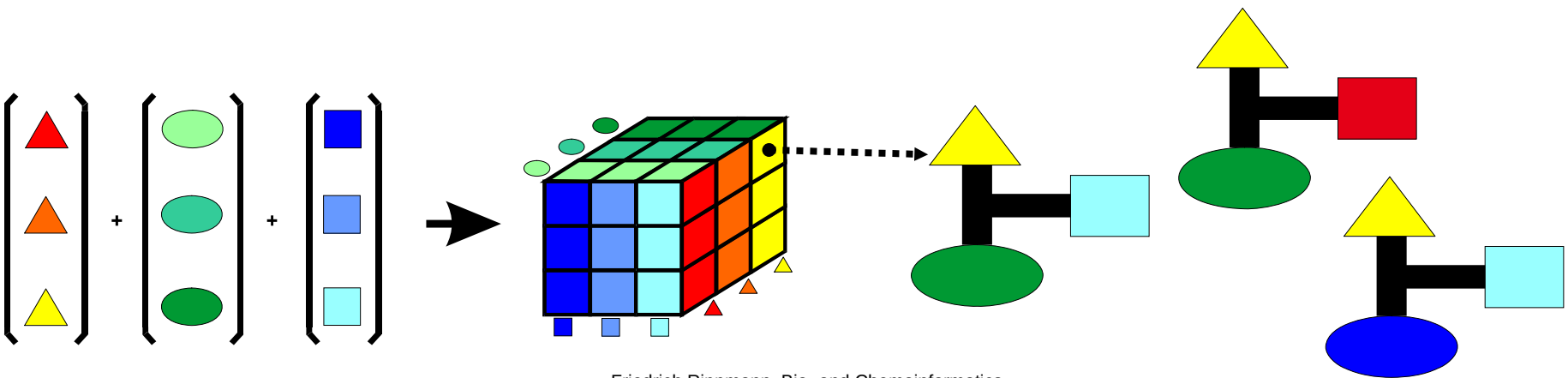
High speed-up by novel algorithm



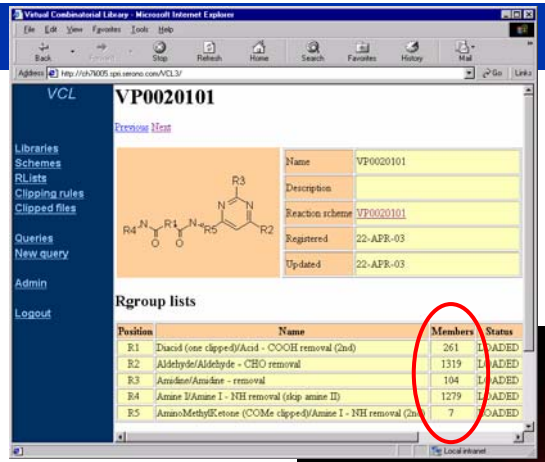
BMBF Project

Application „combinatorial chemistry“

- Possible combinatorial structures need to be generated & filtered
- Typical run time: days to weeks
- Software used: in-house
- Issues: full enumeration prohibitive
- Project impact: medium

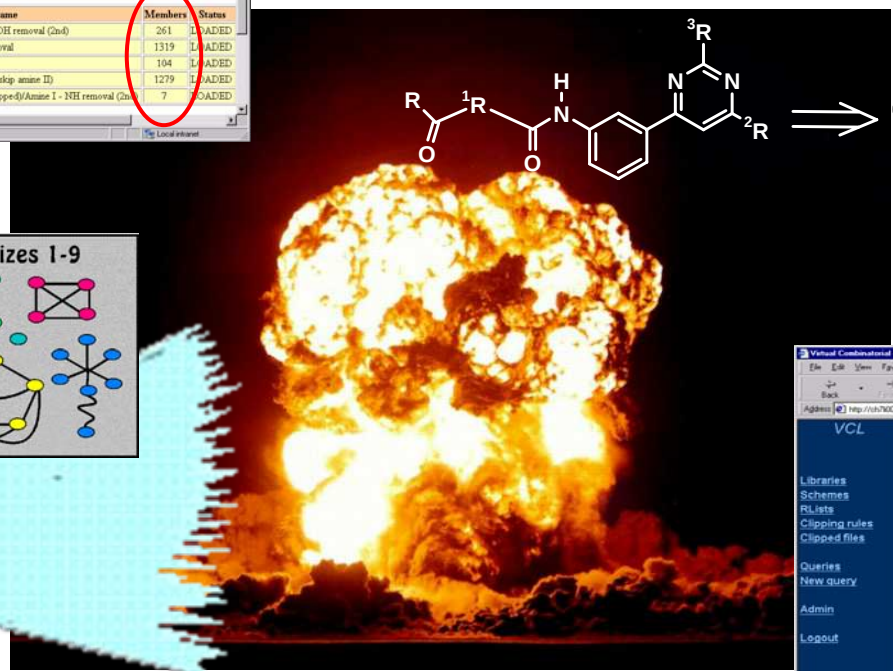


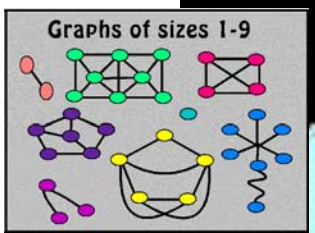
Managing the combinatorial explosion

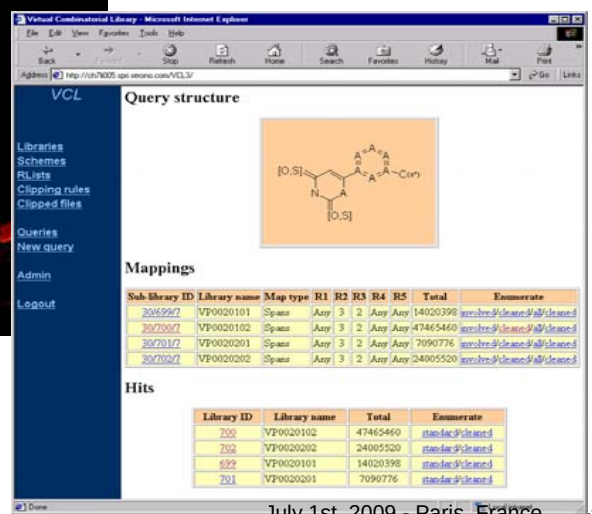


One scaffold and 3000 reagents in 4 groups are enough to generate a library of 320'000'000'000 virtual compounds

In silico screening would take **years** of time and a **petabyte** of space with conventional methods.







Our Virtual Combinatorial Database retrieves all the hits in just a few seconds.

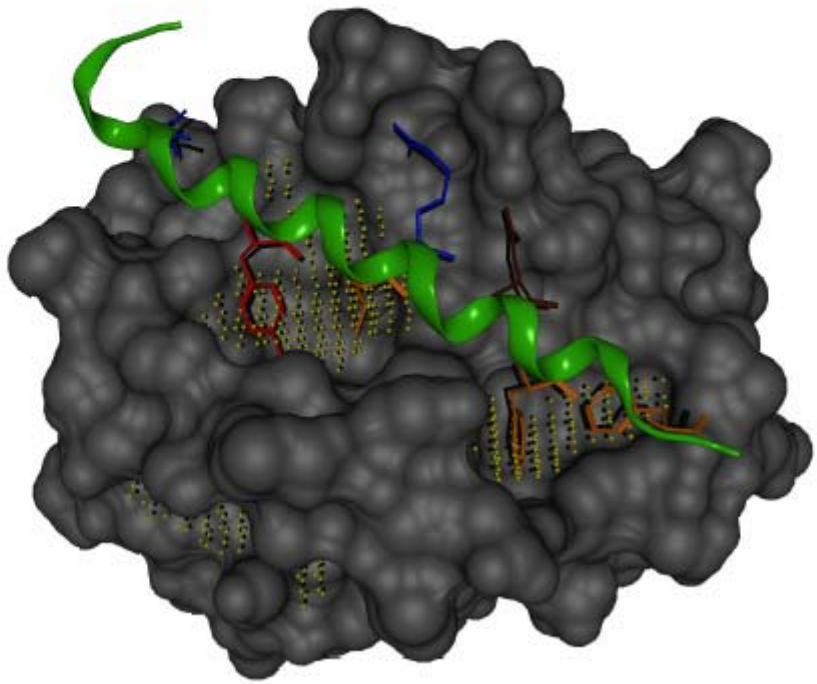
Application „druggability analysis“

- Known protein structures are checked for suitable binding pockets
- Typical run time: days to weeks (followed by long evaluation time)
- Software used: in-house
- Issues: suitable visualization in order to speed-up human analysis
- Project impact: medium

Application „druggability analysis“

Extracted from diploma thesis by Daniela Grimme

Structural analysis



Alanine-scanning mutational analysis:

ΔΔG (kcal /mol)

blue:	2.0 - 3.0
orange:	>3.0 - 4.0
brown:	> 4.0 - 5.0
red:	> 5.0

Identified Pocket Grid Points by the Pocket Analyser:

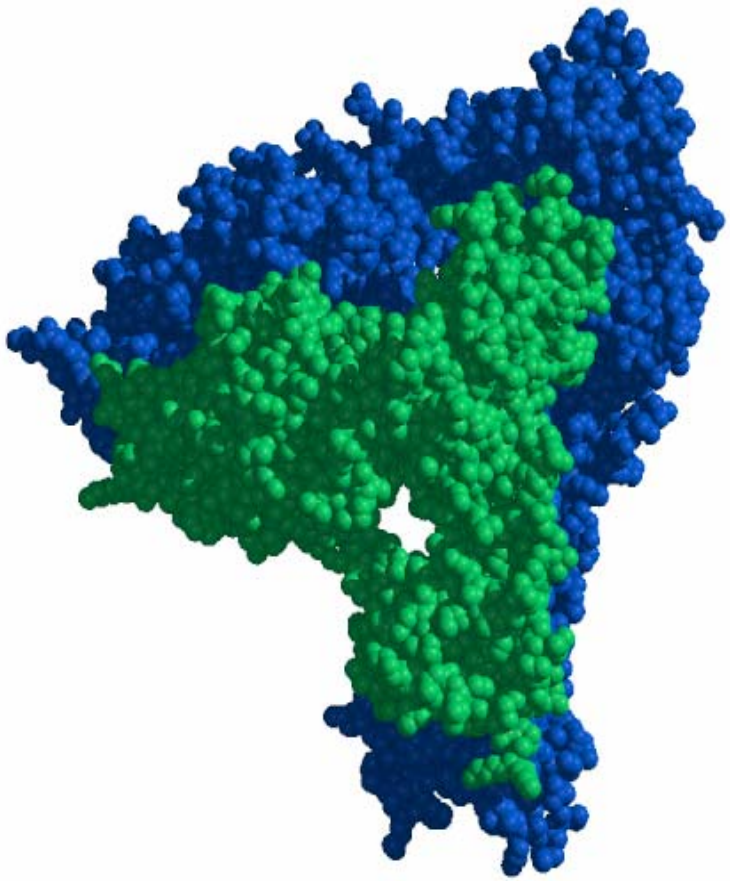
yellow dots

Application „protein dynamics“

- Several applications, e.g. flexibility analysis of proteins (applied to kinases, integrins, antibodies…)
- Typical run time: days
- Software used: AMBER, NMSim/RCNMA, FIRST FRODA, YASARA, …
- Issues: accuracy of parameterization; time scale (pico – nano seconds)
- Project impact: medium

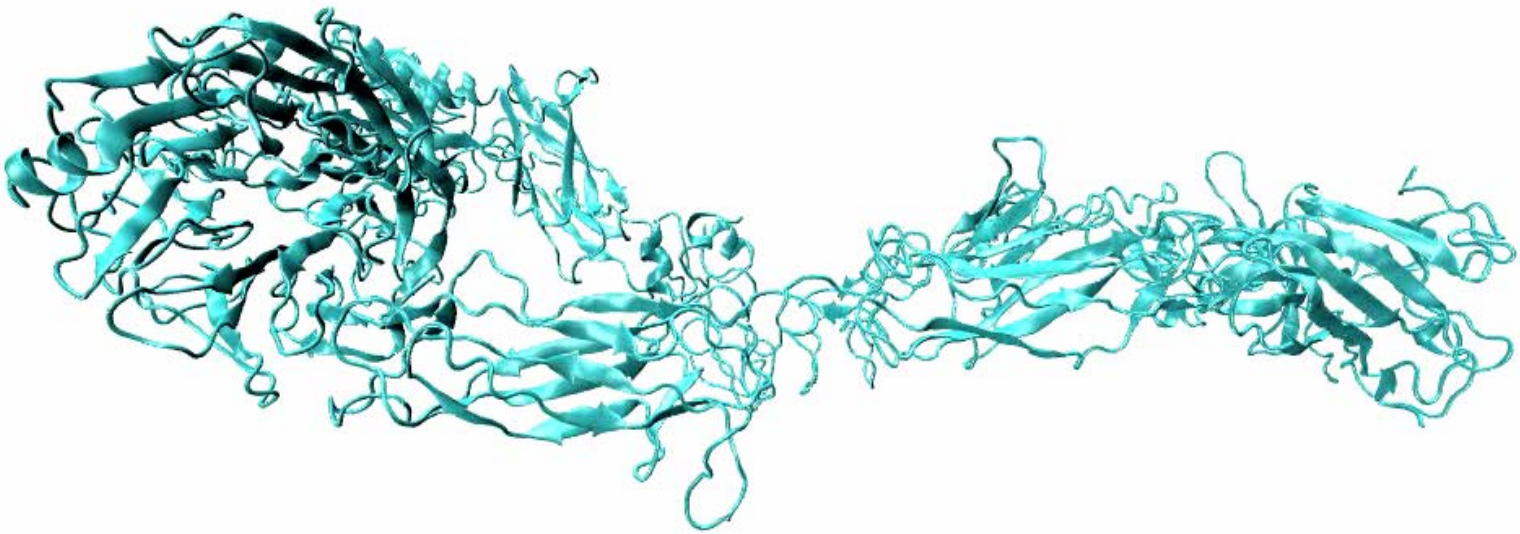
Structure of integrin avb3

Arnaout et al. Science 294 pp. 339 (2001)



Simulation of flexibility

in collaboration with Holger Gohlke



Modelling of antibody flexibility

Simulation done with NMSim/RCNMA, Gohlke et al.

Application „association studies“

- genome-wide association studies: permutations (m markers processed n times, m ~ 1M, n ~ 10000) and genetic interactions (m^2/2 tests)
- Typical run time:
  - Permutation-based FDR computation (10000 permutations) - about one week on 1 CPU
  - Interaction scan (not optimized yet), about 10000 markers - about 12 weeks on 1 CPU
- Software used: in-house
- Project impact: medium

Future applications: Next Generation Sequencing

- "unclassical" HPC problem:
  ~8 processor are enough, but disk speed would not be sufficient to store data as they are generated (there are solutions to this, however)
- Data volumes are too big for transfer to analysis computers (e.g. line speed between Sanger Centre and European Bioinformatics Institute [same campus, physical distance 20m] not sufficient for transfer); internet bandwidth far too small

Nature Biotechnology **25**, 149 (2007)
Published online: 1 February 2007 | doi:10.1038/nbt0207-149

Next-generation sequencing outpaces expectations

Catherine Shaffer[1]

1. Ann Arbor, Michigan

Growing demand in both the research and clinical markets is fueling the development – and funding – of more efficient genomic sequencing methods.

On January 8, Solexa, of Hayward, California, announced the completion of an early-access program evaluating its next-generation Genome Analysis system with customers and reiterated its intention to begin full commercial sales this quarter. Two months earlier, in anticipation of the entry of Solexa's technology and wanting a piece of the emerging market for whole-genome resequencing and analysis, San Diego, California–based microarray maker Illumina announced its intention to acquire the firm in a stock-for-stock transaction valued at around $600 million (*Nat. Biotechnol.* **25**, 10, 2007).

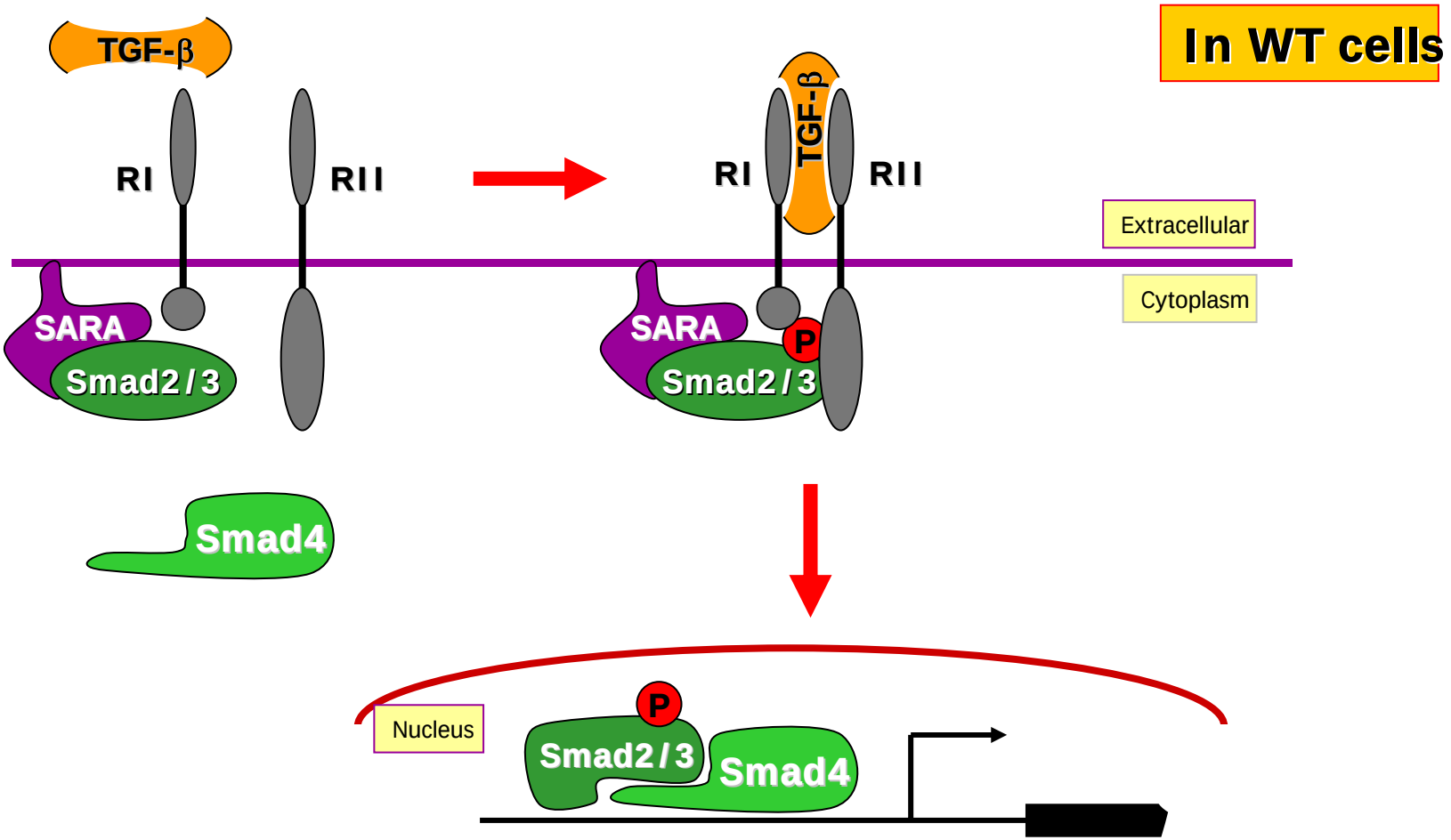
newscom

Next-generation sequencing is already several orders of magnitude more efficient than the Sanger capillary-array electrophoresis (CAE) machines that were the workhorse of the Human Genome Project.

Future applications: Systems Biology

- Some trial applications, but limited acceptance in-house so far
- Simulation of individual pathways, cell populations etc. feasible
- Organ modeling (selected aspects only) becomes possible
- “Virtual human” still some distance away
- Issue: incomplete knowledge, lack of accurate rate constants, protein concentrations, lack of spatial resolution, lack of time resolution

Example: VERY simplified TGF-ß pathway



Current HPC infrastructure (DA only)

Optimized compute server and cluster for analyzing large datasets

Origin 3400

64bit **IRIX** (UNIX of SGI)
- 32 CPUs (**MIPS/R16000**)
- 32 GByte *Shared* Memory

Altix 3700

64bit **Linux**
- 32 CPUs (INTeL **Itanium**)
- 256 GByte *Shared* Memory

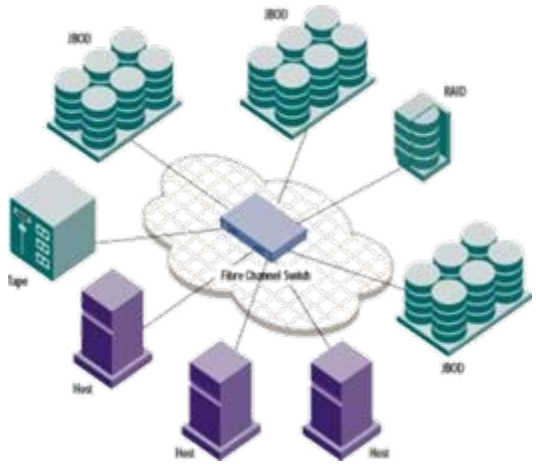
For MEDIUM number of LARGE problems

Linux-Cluster

64bit **Linux** (Cluster)
- 94x 4 CPU-Cores (Xeon)
- 94x 8 GByte *distinct* Memory

For LARGE number of SMALL problems

High-end graphics workstations with 3D stereo support



7.5 TByte Shared Storage for biological and chemical databases and analysis results

IT infrastructure optimized for high performance and high throughput computing, using a large pool of sophisticated scientific software

Future challenges

- Compute power less an issue, but DATA deluge (e.g. personal genomes; quote from director of German Cancer Research Center: “In five years we will sequence all tumors of all our patients”
  - Storage
  - Bandwidth locally/Internet
- Power consumption may become limiting (1 entire nuclear power plant is needed just to serve German HPC centers)

The end…

…and finally we have calculated that our simulations have added 0.5 degrees to Global Warming…

Thanks to

- Reinhard Schneider, Chris Sander, Alexander Reinefeld, Willie Taylor, Andras Aszodi, Holger Gohlke
- Jacques Barbanton, Thierry Convard
- Massimo de Francesco, Jerome Wojcik
- Christian Griebel, Anja von Heydebreck, Oliver Karch, Michael Krug, Mireille Krier, Daniela Grimme