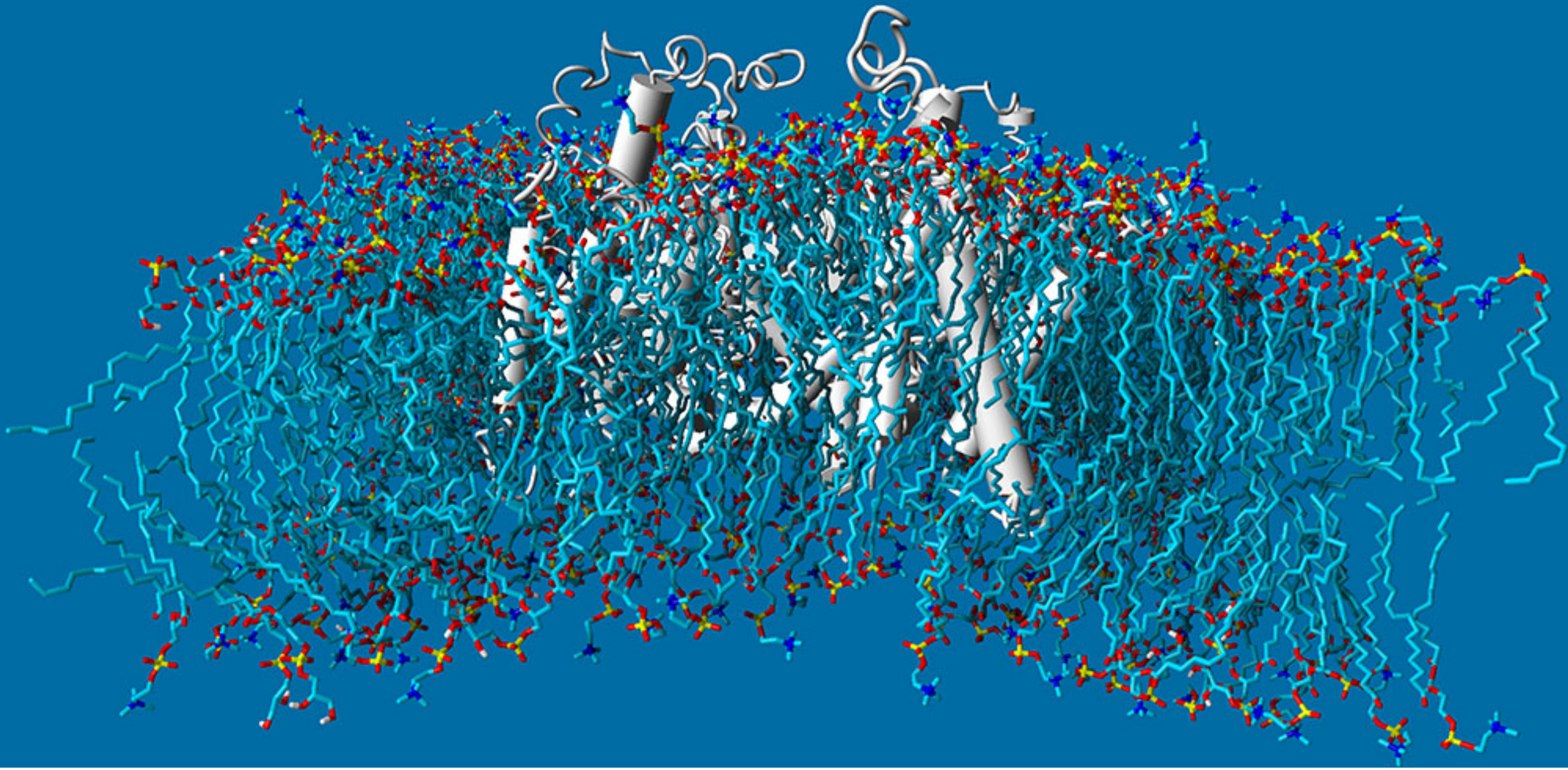
ADDED VALUE GIVEN BY LONG-TERM STORAGE FOR COMPUTATIONAL SCIENCES

Ilpo Vattulainen

Tampere Univ of Tech & MEMPHYS-SDU Odense

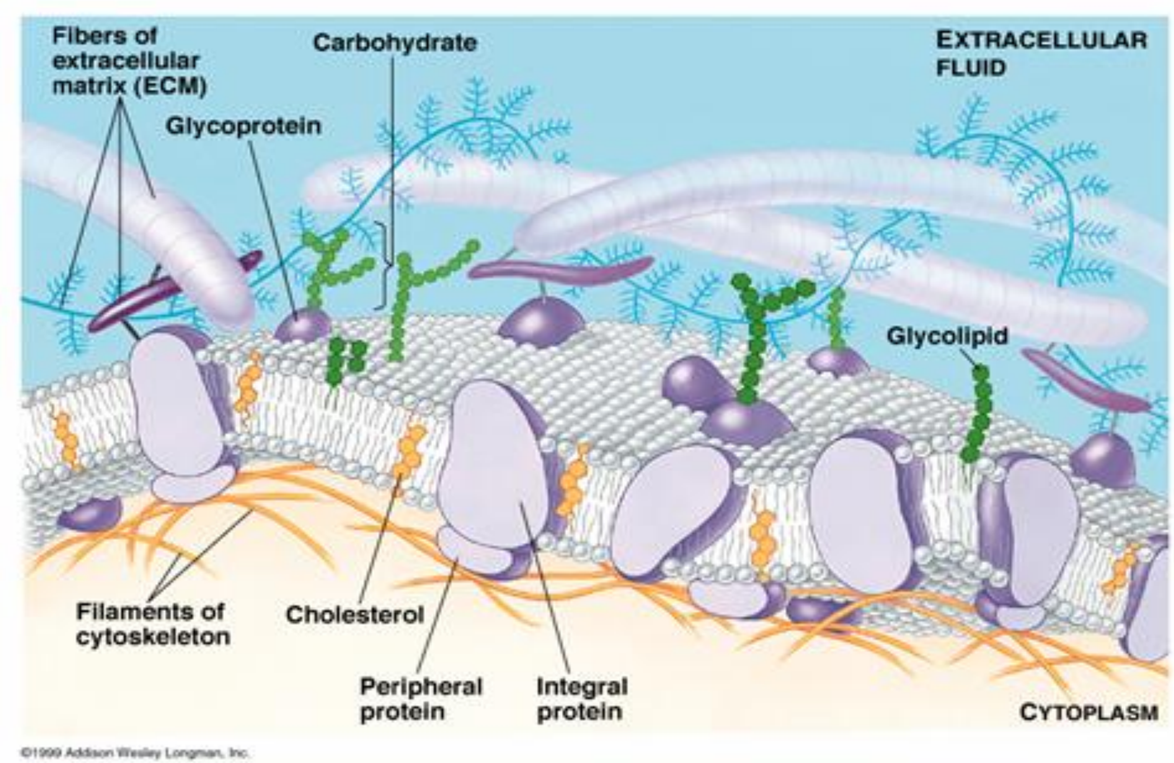
Biological Physics Group – ERC Advanced Grant for 2012-2017

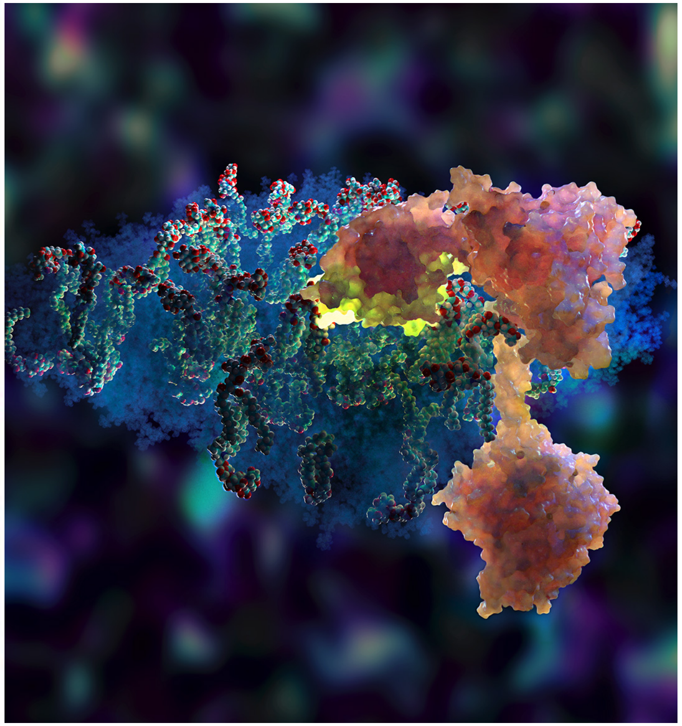


Case in Brief

- Molecular simulations generating 10^x Terabytes of data
- How to store the data for analysis over a period of 3-10-20 years?

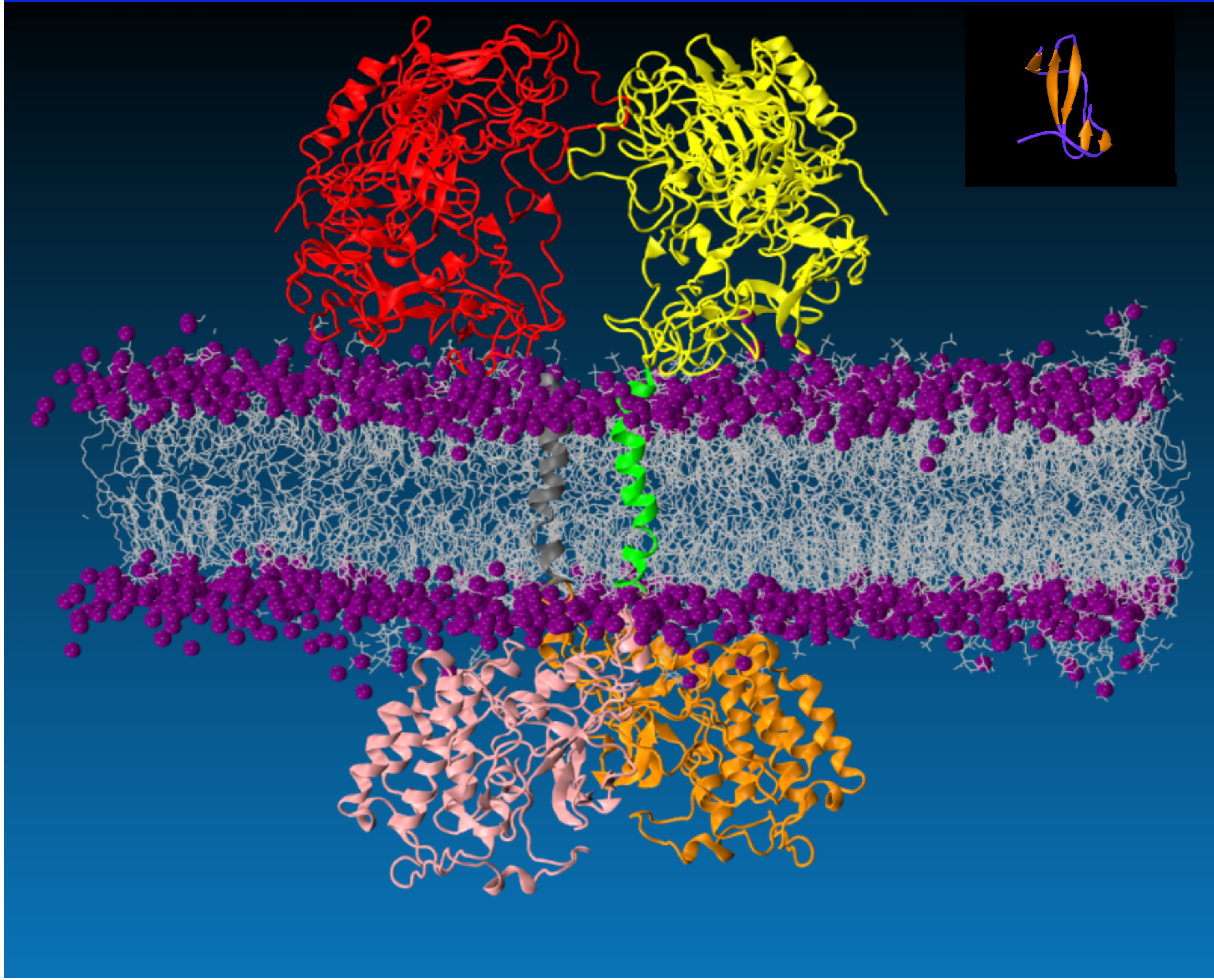
Biological Context: Proteins and Other Receptors





Nanoscale engines in cells –
Receptor function bridged to conformation

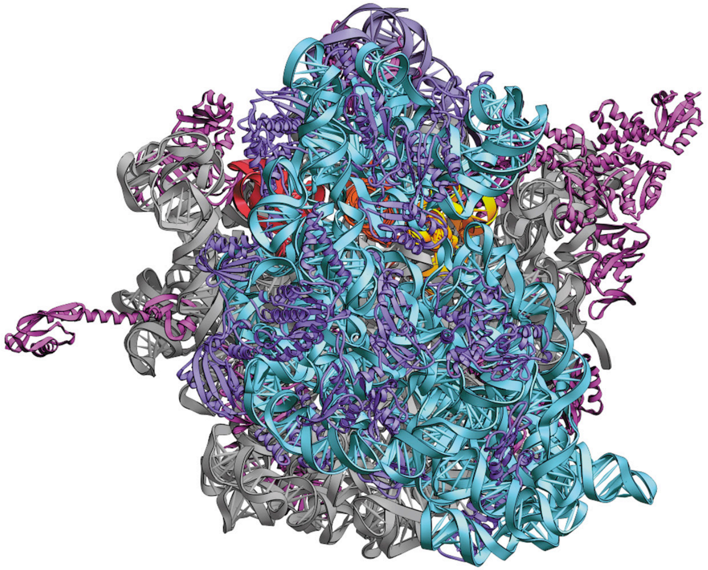
Membrane Receptors Targeted by Drugs



- Example: GPCRs
- Target of ~50% of drug development
- Annual revenue above USD 65 billion
- The case shown for epidermal growth factor receptor (EGFR) dimerization/ activation

Molecular Simulations

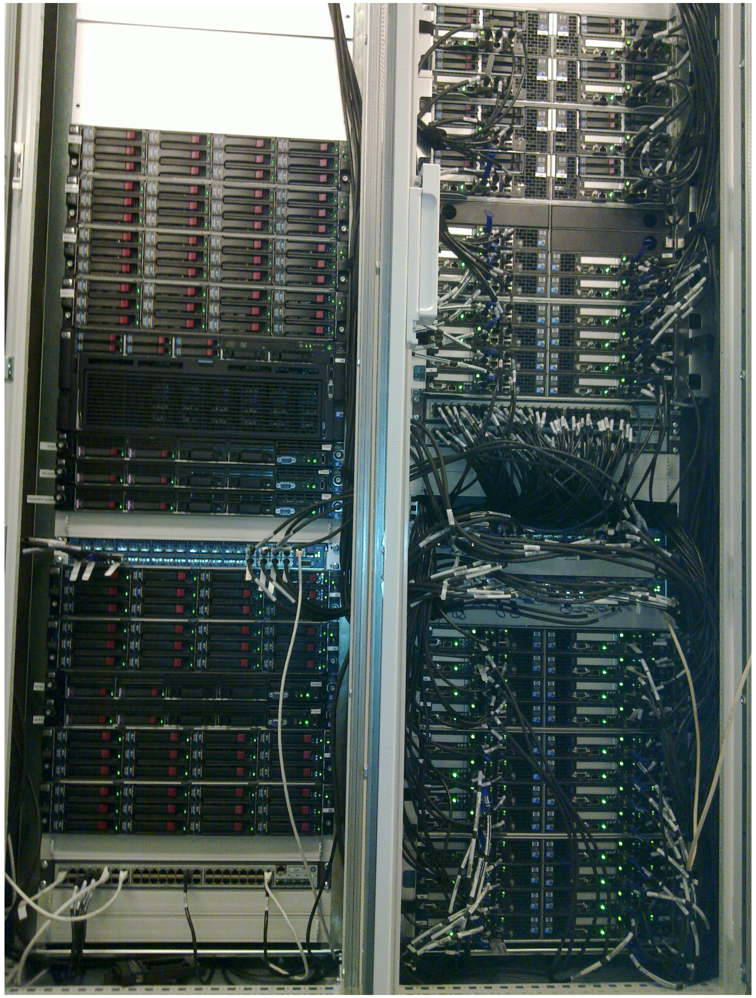
$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i \qquad \boldsymbol{f}_i = -\frac{\partial}{\partial \boldsymbol{r}_i}\mathcal{U}$$



$$\mathcal{U}_{\text{non-bonded}} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] + \frac{Q_1 Q_2}{4\pi\epsilon_0 r}$$

$$\mathcal{U}_{\text{intramolecular}} = \frac{1}{2}\sum_{\text{bonds}} k^{r}_{ij}\left(r_{ij} - r_{\text{eq}}\right)^2 + \frac{1}{2}\sum_{\substack{\text{bend}\\ \text{angles}}} k^{\theta}_{ijk}\left(\theta_{ijk} - \theta_{\text{eq}}\right)^2 + \frac{1}{2}\sum_{\substack{\text{torsion}\\ \text{angles}}}\sum_{m} k^{\phi,m}_{ijkl}\left(1 + \cos(m\phi_{ijkl} - \gamma_m)\right)$$

Resources: Computing



Tampere Center for Scientific Computing

- directed by Vattulainen

Main <u>computer nodes</u> based on the *HP ProLiant SL6500 Scalable System* product family (installed in Dec 2011)

- HP ProLiant SL390s G7 1U half width server
- 2 * Intel 6-core Xeon X5650 CPU
- 48-96 GB memory (4GB-8GB/core)
- HP ProLiant s6500 4U Chassis
- 2000 CPU cores (by end of 2013)

Resources: Computing

CSC – IT Centre for Science

- Cray XC30 + other machines
- About 12,000 cores
- Upgraded in Dec 2012

Also, access to various other supercomputing centres:

- Tier-0 resources (PRACE): 60,000,000 core-hours granted in Feb 2013
- Tier-1 resources in DECI/PRACE (EU FP7)
- Jugine in Julich
- HorseShoe in Odense, Denmark
- SharcNet in Canada
- Etc.

We use ~10,000 core-years of computing time in 2013.

Amount of Data We Get Today

- A typical simulation for ~200,000 atoms over 1 microsecond: 200 GB of data
- About 10 simulations per project: 2 Tb
- About 40 members in the team, each with a project. Total data: ~100 Tb per year
- Data storage: On local computers, external hard disks (CSC quotas overused by almost all of our people)
- State of the art simulations require even larger data storage resources: The PRACE project alone (60,000,000 core-hours) will generate ~10 Tb of data.

How to Deal with Massive Data

Short-term storage (~1-3 years)

Multiple-backup principle for trajectories

- ***Local*** **disks in large servers up to multiple Tb's (expensive, safe) – long-term storage**
- ***Local in-group*** **internal and external hard disks (cheap, vulnerable) – needed for analysis**
- ***National*** **backups: CSC (archive; safe, limited by CSC resources, not fast) – long-term storage**

How to Deal with Massive Data

Long-term storage (~10 years)

Backups for the primary simulation files of systems that have been simulated, stored as a database for

- ***Starting and end configuration files***
- ***Force field***
- ***Simulation/run files***
- ***Article versions prepared (PDF, doc, etc.)***
- ***Analysis codes***

Added Value of Long-Term Storage?

Primary focus of long-term storage:

- Only limited primary data is always stored permanently: files needed to repeat the simulations
- Given these, if needed, the simulations can (usually) be repeated with minor computing resources 4-5 years later

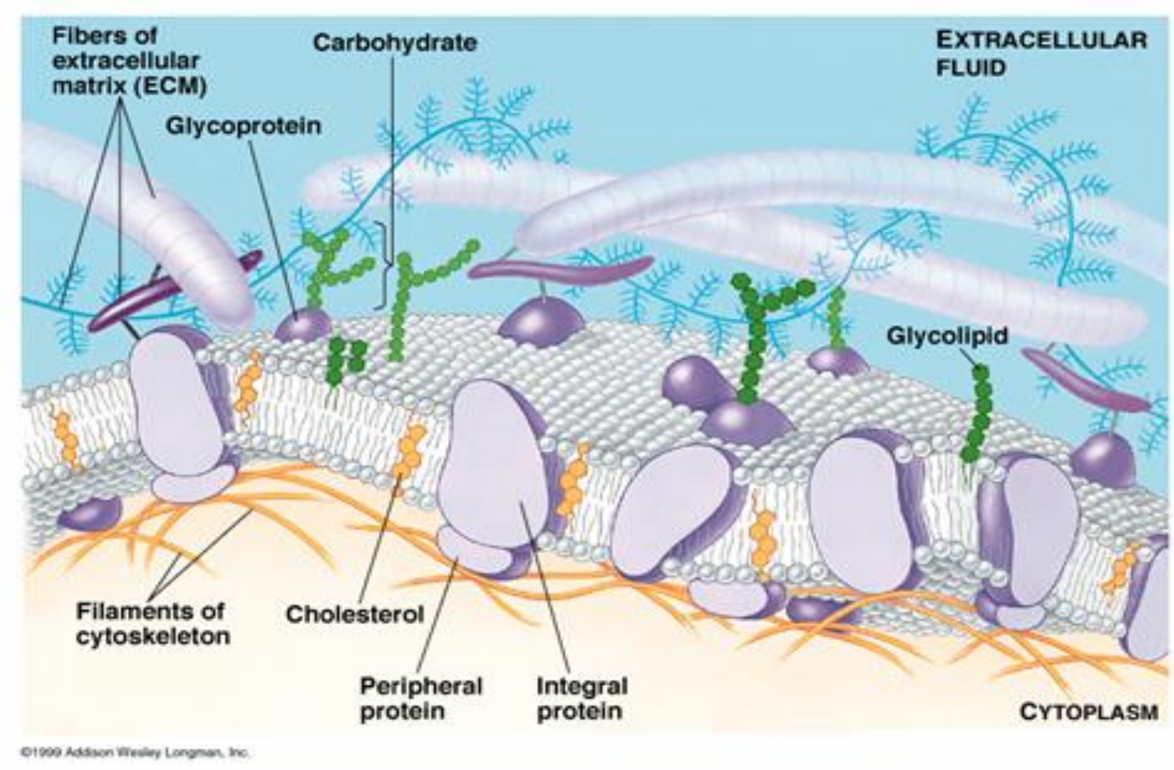
However:

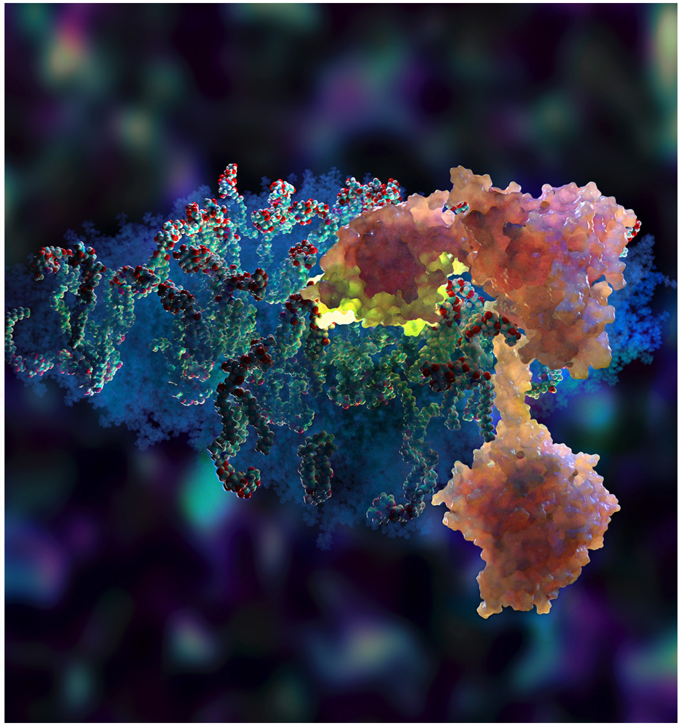
- New versions of simulation packages are occasionally not compatible with older versions, implying that the simulations cannot repeated identically even if all the input files are available
- Occasional need to reconsider older results to consider the quality of the models used
- Based on new exptl evidence, new analysis of older data would be preferable
- Development of theoretical descriptions requires older data to be at hand

Secondary focus of long-term storage:

- Resources allowing, we store all the simulation data we have

After All There is the Biological Context





Understanding receptor function allows design of new means to control the function for better health

Thank you

Acknowledgments:

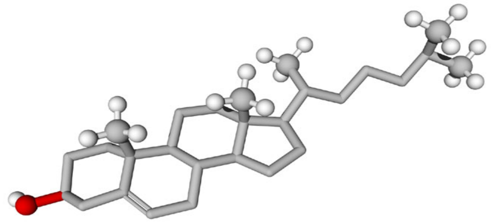
- Academy of Finland
- Center of Excellence funding
- ERC, EU FP7
- ESF, Nordita
- Bioindustry, Tekes
- Finnish Foundations
- CSC

Collaborators (experiments, examples):

Elina Ikonen (Biomedicum, Hki)
Kai Simons (MPI Dresden)
Christian Eggeling (Göttingen)
Petri Kovanen (Wihuri, Hki)
Amy Rowat (Harvard)
Peter Westh (Roskilde)
Juha Holopainen (HUS, Hki)
Susanne Wiedmer (Hki)
Michael R. Morrow (Newfoundland)
Filip Tuomisto (Aalto)

Collaborators (theory, examples):

Mikko Karttunen (Ontario)
Roland Faller (UC Davis)
Tapio Ala-Nissilä (Aalto)
Adam Foster (TUT)
Matej Oresic et al. (VTT)
P. Capkova (Prague)
Aatto Laaksonen et al. (Stockholm)
Erik Lindahl (Stockholm)
Alex Bunker (Viikki/Drug Design)
SJ Marrink (Groningen)
AA Gurtovenko (UK)



Networks:

SimBioMa (ESF), MOLSIMU (COST), Nordita, Graduate Schools, etc.