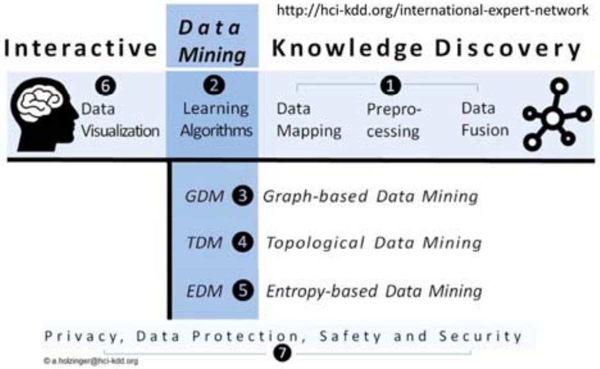
Andreas Holzinger
VO 709.049 Medical Informatics
30.11.2016 11:15-12:45

Lecture 06 Probabilistic Graphical Models II: From Bayesian Networks to Graph Bandits

a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data

ML needs a concerted effort fostering integrated research


http://hci-kdd.org/international-expert-network
Interactive
Data Mining
Knowledge Discovery
6 Data Visualization
2 Learning Algorithms
1 Data Mapping
Prepro-cessing
Data Fusion
GDM 3 Graph-based Data Mining
TDM 4 Topological Data Mining
EDM 5 Entropy-based Data Mining
Privacy, Data Protection, Safety and Security
7
© a.holzinger@hci-kdd.org

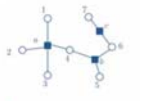
Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Agenda for today

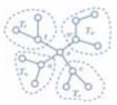
- 00 Reflection – follow-up from last lecture
- 01 Graphical Models and Decision Making
- 02 Bayesian Networks
- 03 Machine Learning on Graphs
- 04 Little Excursus: What is similarity?
- 05 Probabilistic Topic Models
- 06 Graph Bandits (a very hot topic!)

00 Reflection

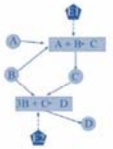
Quiz



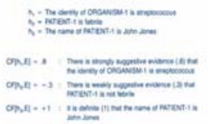
1



2



3



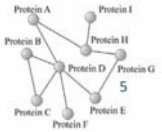
h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

CF[h_1,E] = .8 : There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

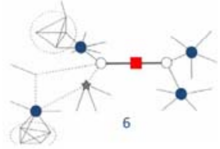
CF[h_2,E] = −.3 : There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

CF[h_3,E] = +1 : It is definite (1) that the name of PATIENT-1 is John Jones

4

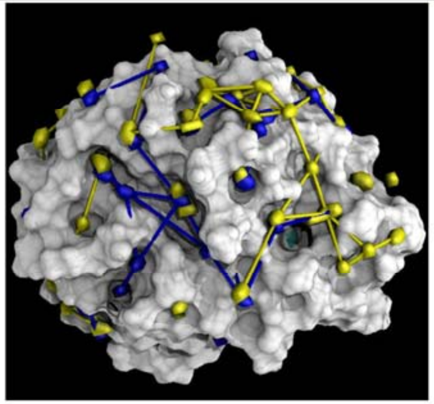


5



6

Quiz



http://sbcb.bioch.ox.ac.uk/users/oliver/software/

Graph

Model

$\mathcal{M}$

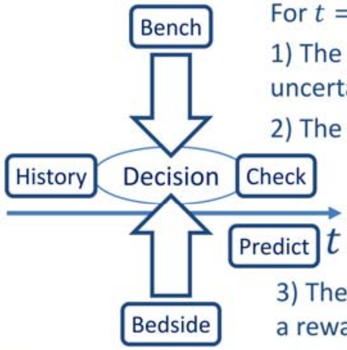
01 Graphical Models and Decision Making

Data

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}_{i=1}^{N}$$

Decision Making: Learn good policy for selecting actions

Goal: Learn an **optimal policy** for selecting best actions within a given **context**



For $t = 1, \ldots, T$

1) The world produces an uncertain “context” $x_t \in X$

2) The learner selects an action $a_t \in \{1, \ldots, K\}$

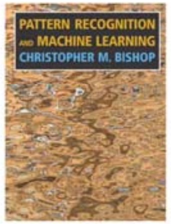
3) The world reacts with a reward $r_t(a_t) \in [0,1]$

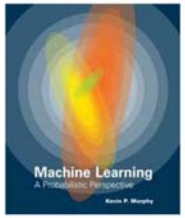
Key Challenges

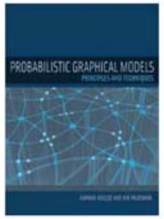
- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- When we have big data but little knowledge automatic ML can help to gain insight:
- **Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**
- If we have little data and deal with NP-hard problems we still need the human-in-the-loop!

Example: Why is prediction in proteins so important for us?

Recommended background reading





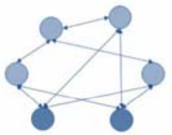


Bishop, C. M. 2007. Pattern Recognition and Machine Learning, Heidelberg, Springer. Chapter 8 on graphical models openly available: http://research.microsoft.com/en-us/um/people/cmbishop/prml/

Murphy, K. P. 2012. Machine learning: a probabilistic perspective, MIT press. Chapter 26 (pp. 907) – Graphical model structure learning

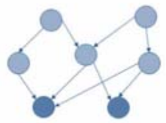
Koller, D. & Friedman, N. 2009. Probabilistic graphical models: principles and techniques, MIT press.

Three types of Probabilistic Graphical Models



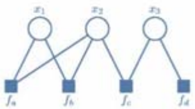
Undirected: Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

$$P(X)=\frac{1}{Z}\exp\left(\sum_{ij}W_{ij}\,x_i x_j+\sum_i x_i b_i\right)$$



Directed: Bayes Nets, useful for designing models (Details: Murphy 10)

$$p(\mathbf{x})=\prod_{k=1}^{K}p(x_k|\mathrm{pa}_k)$$



Factored: useful for inference/learning

$$p(\mathbf{x})=\prod_s f_s(\mathbf{x}_s)$$

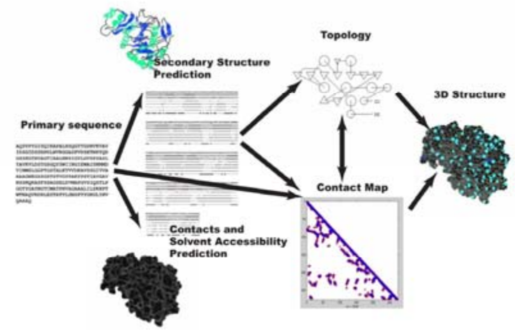
Factor Graphs – learning at scale

- What is the advantage of factor graphs?

	Dependency	Efficient Inference	Usage
Bayesian Networks	Yes	Somewhat	Ancestral Generative Process
Markov Networks	Yes	No	Local Couplings and Potentials
Factor Graphs	No	Yes	Efficient, distributed inference

Table credit to Ralf Herbrich, Amazon

From structure to function prediction



Secondary Structure Prediction
Topology
3D Structure
Primary sequence
Contact Map
Contacts and Solvent Accessibility Prediction

Baldi, P. & Pollastri, G. 2003. The principled design of large-scale recursive neural network architectures--dag-rnns and the protein structure prediction problem. The Journal of Machine Learning Research, 4, 575-602.

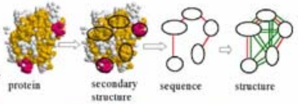
Protein Network Inference

- Hypothesis: most biological functions involve the interactions between many proteins, and the complexity of living systems arises as a result of such interactions.
- In this context, the problem of inferring a global protein network for a given organism,
- - using all (genomic) data of the organism,
- is one of the main challenges in computational biology

Yamanishi, Y., Vert, J.-P. & Kanehisa, M. 2004. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics, 20, (suppl 1), i363-i370.

Problem: Is Graph Isomorphism NP-complete ?

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S., Smola, A. J. & Kriegel, H.-P. 2005. Protein function prediction via graph kernels. Bioinformatics, 21, (suppl 1), i47-i56.



- Important for health informatics: Discovering relationships between biological components
- Unsolved problem in computer science:
- Can the graph isomorphism problem be solved in polynomial time?
  - So far, no polynomial time algorithm is known.
  - It is also not known if it is NP-complete
  - We know that subgraph-isomorphism is NP-complete

Example: Protein Network Inference

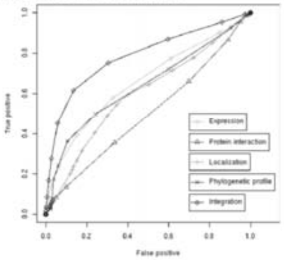
BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i363–i370
DOI: 10.1093/bioinformatics/bth910

Protein network inference from multiple genomic data: a supervised approach

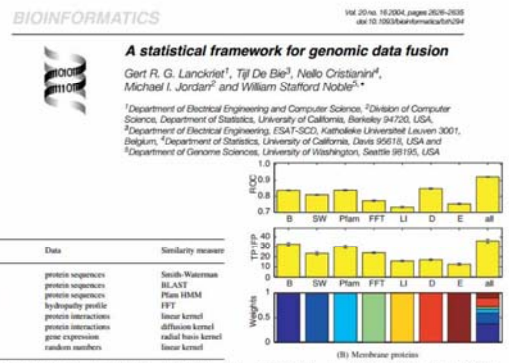
Y. Yamanishi[1,], J.-P. Vert[2] and M. Kanehisa[1]*

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan and [2] Computational Biology group, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France



K_{exp} (Expression)
K_{ppi} (Protein interaction)
K_{loc} (Localization)
K_{phy} (Phylogenetic profile)
$K_{\text{exp}}+K_{\text{ppi}}+K_{\text{loc}}+K_{\text{phy}}$ (Integration)

Example: Data fusion and Protein Annotation


BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2626–2635
doi:10.1093/bioinformatics/bth294

A statistical framework for genomic data fusion

Gert R. G. Lanckriet[1], Tijl De Bie[3], Nello Cristianini[4], Michael I. Jordan[2] and William Stafford Noble[5,*]

[1]Department of Electrical Engineering and Computer Science, [2]Division of Computer Science, Department of Statistics, University of California, Berkeley 94720, USA, [3]Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven 3001, Belgium, [4]Department of Statistics, University of California, Davis 95618, USA and [5]Department of Genome Sciences, University of Washington, Seattle 98195, USA

Kernel	Data	Similarity measure
K_{SW}	protein sequences	Smith-Waterman
K_B	protein sequences	BLAST
K_{Pfam}	protein sequences	Pfam HMM
K_{FFT}	hydropathy profile	FFT
K_{LI}	protein interactions	linear kernel
K_D	protein interactions	diffusion kernel
K_E	gene expression	radial basis kernel
K_{RND}	random numbers	linear kernel

(B) Membrane proteins

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. 2004. A statistical framework for genomic data fusion. Bioinformatics, 20, (16), 2626-2635.

02 Bayesian Networks "Bayes' Nets"

Bayesian Network (BN) - Definition

- is a **probabilistic model**, consisting of two parts:
- 1) a dependency structure and
- 2) local probability models.

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, Pa(x_i))$$

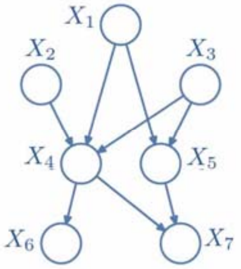
Where $Pa(x_i)$ are the parents of x_i

BN inherently model the **uncertainty in the data.** They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources.

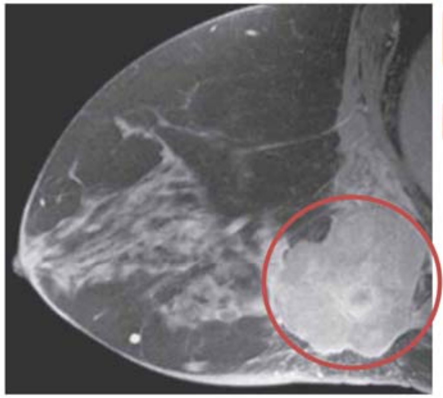
Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Example: Directed Bayesian Network with 7 nodes

$$p(X_1, \dots, X_7) = p(X_1)p(X_2)p(X_3)p(X_4|X_1,X_2,X_3)\cdot p(X_5|X_1,X_3)p(X_6|X_4)p(X_7|X_4,X_5)$$



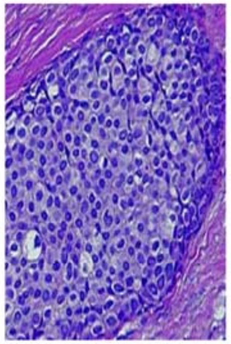
Clinical Case Example



Overmoyer, B. A., Lee, J. M. & Lerwill, M. F. (2011) Case 17-2011 A 49-Year-Old Woman with a Mass in the Breast and Overlying Skin Changes. *New England Journal of Medicine, 364, 23, 2246-2254.*

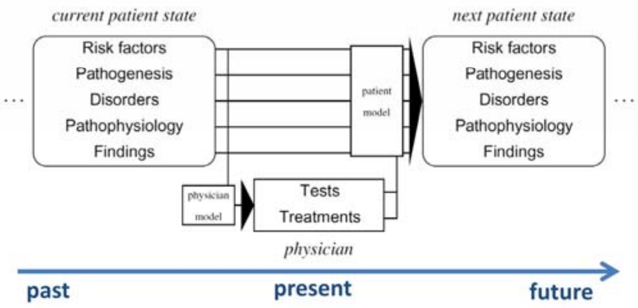
Important in Clinical practice -> prognosis !

- = the prediction of the future course of a disease conditional on the patient's history and a projected treatment strategy
- Danger: probable Information !
- Therefore valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality of-life decisions



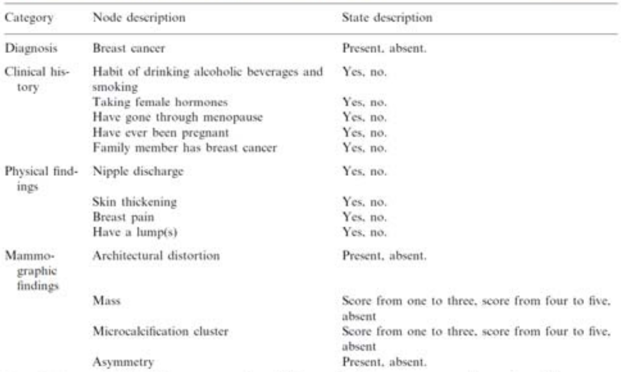
Knaus, W. A., Wagner, D. P. & Lynn, J. (1991) Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science, 254, 5030, 389.*

Predicting the future on past data and present status



van Gerven, M. A. J., Taal, B. G. & Lucas, P. J. F. (2008) Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics, 41, 4, 515-529.*

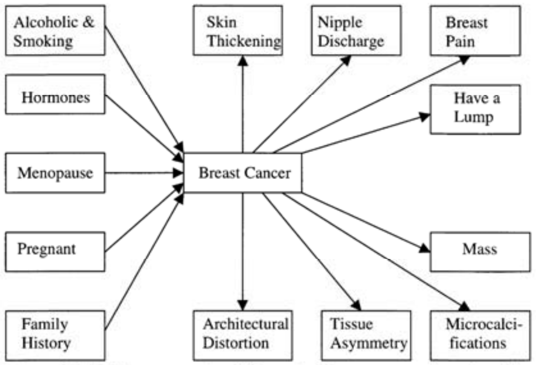
Example: Breast cancer - Probability Table



Category	Node description	State description
Diagnosis	Breast cancer	Present, absent.
Clinical history	Habit of drinking alcoholic beverages and smoking	Yes, no.
	Taking female hormones	Yes, no.
	Have gone through menopause	Yes, no.
	Have ever been pregnant	Yes, no.
	Family member has breast cancer	Yes, no.
Physical findings	Nipple discharge	Yes, no.
	Skin thickening	Yes, no.
	Breast pain	Yes, no.
	Have a lump(s)	Yes, no.
Mammographic findings	Architectural distortion	Present, absent.
	Mass	Score from one to three, score from four to five, absent
	Microcalcification cluster	Score from one to three, score from four to five, absent
	Asymmetry	Present, absent.

Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*

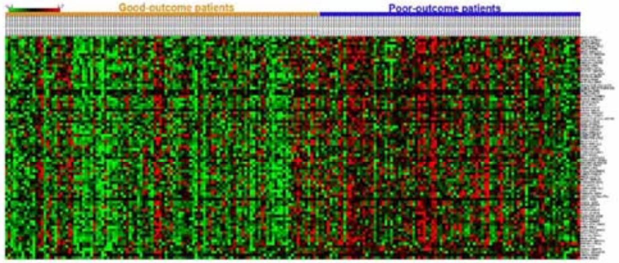
Breast cancer – big picture – state of 1999



Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*

10 years later: Integration of microarray data

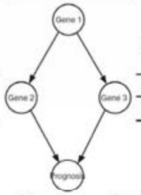
- Integrating microarray data from multiple studies to increase sample size;
- = approach to the development of more robust prognostic tests



Xu, L., Tan, A., Winslow, R. & Geman, D. (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics, 9, 1, 125-139.*

Example: Bayes Net with four binary variables

Gene 1	
P(on)	0.8
P(off)	0.2



Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

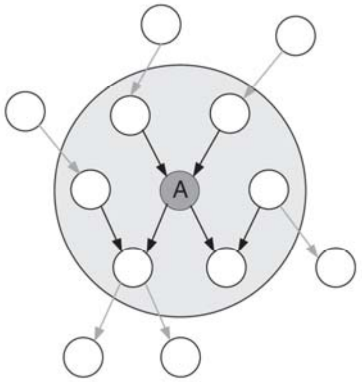
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 2 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

Concept Markov-Blanket

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*



Dependency Structure -> first step (1/2)

- First the structure is learned using a search strategy.
- Since the number of possible structures increases super exponentially with the number of variables,
- the well-known greedy search algorithm K2 can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]$$

N_{ijk} … number of cases in the data set D
having variable i in state k associated with the j-th instantiation of its parents in current structure S.
n is the total number of variables.

Dependency Structure – first step (2/2)

- Next, N_{ij} is calculated by summing over all states of a variable:
- $N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \cdot N'_{ijk}$ and N'_{ij} have similar meanings but refer to prior knowledge for the parameters.
- When no knowledge is available they are estimated using $N_{ijk} = N/(r_i q_i)$
- with N the equivalent sample size,
- r_i the number of states of variable i and
- q_i the number of instantiations of the parents of variable i.
- $\Gamma(.)$ corresponds to the gamma distribution.
- Finally $p(S)$ is the prior probability of the structure.
- $p(S)$ is calculated by:
- $p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{p_i} p(l_i \longrightarrow x_i) \prod_{m_i=1}^{o_i} p(m_i x_i)$
- with p_i the number of parents of variable x_i and o_i all the variables that are not a parent of x_i.
- Next, $p(a \longrightarrow b)$ is the probability that there is an edge from a to b while $p(ab)$ is the inverse, i.e. the probability that there is no edge from a to b

Parameter learning -> second step

- Estimating the parameters of the local probability models corresponding with the dependency structure.
- CPTs are used to model these local probability models.
- For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters.
- Each set of parameters was given a uniform Dirichlet prior:

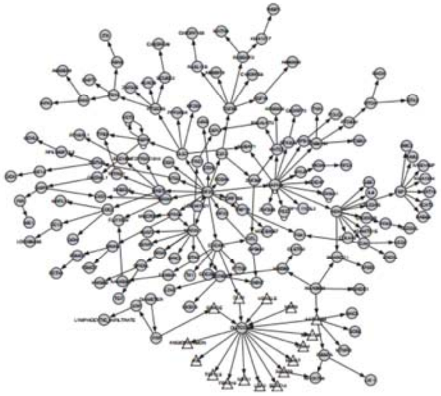
$$\mathrm{p}(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, \dots, N'_{ijk}, \dots, N'_{ijr_i})$$

Note: With θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, \dots, N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

$$\mathrm{p}(\theta_{ij}|D,S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, \dots, N'_{ijk} + N_{ijk}, \dots, N'_{ijr_i} + N_{ijr_i})$$

with N_{ijk} defined as before.

Predicting the prognosis of breast cancer (integrated a.)



Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

Inference in Bayes Nets is intractable (NP-complete!)

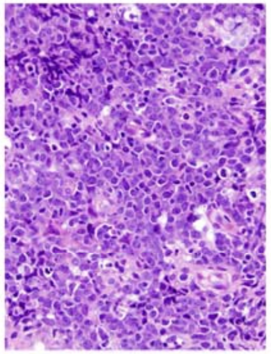
- For certain cases it is tractable if:
  - Just one variable is unobserved
  - We have singly connected graphs (no undirected loops -> belief propagation)
  - Assigning probability to fully observed set of variables
- Possibility: Monte Carlo Methods (generate many samples according to the Bayes Net distribution and then count the results)
- Otherwise: approximate solutions, NOTE:

Sometimes it is better to have an approximate solution to a complex problem – than a perfect solution to a simplified problem

3) Machine Learning on Graphs

Example: Lymphoma is the most common blood cancer

The two main forms of lymphoma are Hodgkin lymphoma and non-Hodgkin lymphoma (NHL). Lymphoma occurs when cells of the immune system called lymphocytes, a type of white blood cell, grow and multiply uncontrollably. Cancerous lymphocytes can travel to many parts of the body, including the lymph nodes, spleen, bone marrow, blood, or other organs, and form a mass called a tumor. The body has two main types of lymphocytes that can develop into lymphomas: B-lymphocytes (B-cells) and T-lymphocytes (T-cells).

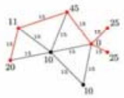


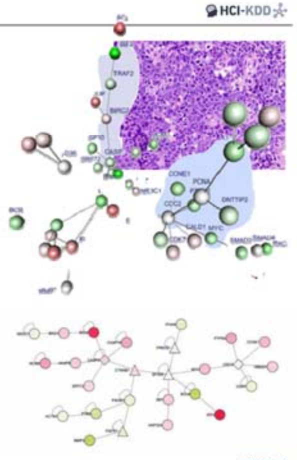
www.lymphoma.org

http://imagebank.hematology.org/

ML tasks on graphs

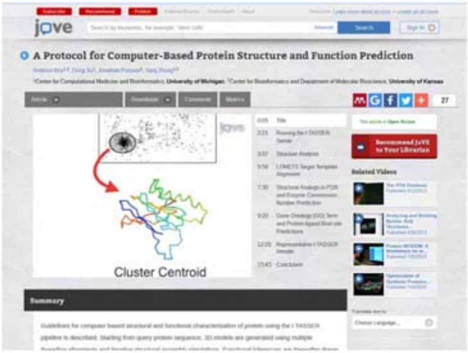
- **Discover** unexplored interactions in PPI-networks and gene regulatory networks
- **Learn** the structure
- **Reconstruct** the structure





Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Müller, T. 2008. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics, 24, (13), i223-i231.

From structure to function



A Protocol for Computer-Based Protein Structure and Function Prediction

http://www.jove.com/video/3259/a-protocol-for-computer-based-protein-structure-function

Interesting: Hubs tend to link to small degree nodes

Nodes: proteins

Links: physical interactions (binding)

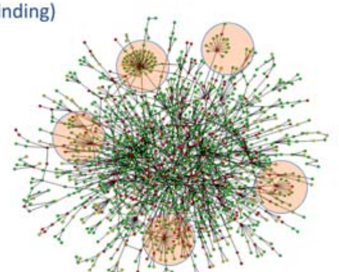
Puzzling pattern:
Hubs tend to link to small degree nodes.
Why is this puzzling?

In a random network, the probability that a node with degree *k* links to a node with degree *k'* is:

$$p_{kk'} = \frac{kk'}{2L}$$

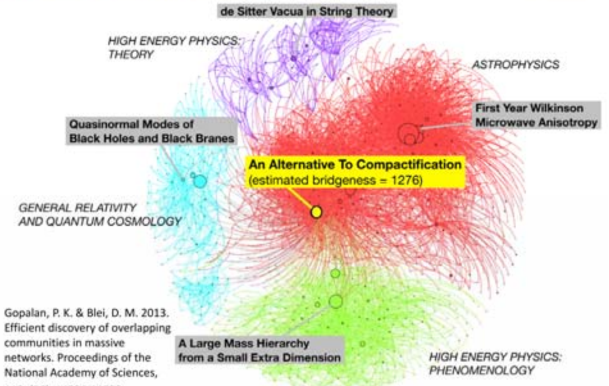
k=50, k'=13, N=1,458, L=1746

$p_{50,13} = 0.15$ $p_{2,1} = 0.0004$



Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. 2001. Lethality and centrality in protein networks. Nature, 411, (6833), 41-42.

Example: Subgraph Discovery

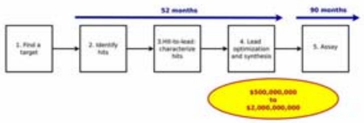


de Sitter Vacua in String Theory
HIGH ENERGY PHYSICS: THEORY
ASTROPHYSICS
Quasinormal Modes of Black Holes and Black Branes
First Year Wilkinson Microwave Anisotropy
An Alternative To Compactification (estimated bridgeness = 1276)
GENERAL RELATIVITY AND QUANTUM COSMOLOGY
A Large Mass Hierarchy from a Small Extra Dimension
HIGH ENERGY PHYSICS: PHENOMENOLOGY

Gopalan, P. K. & Blei, D. M. 2013. Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110, (36), 14534-14539.

Why do we want to apply ML to graphs

- A) Discovery of unexplored interactions
- B) Learning and Predicting the structure
- C) Reconstructing the structure
- Which joint probability distributions does a graphical model represent?
- How can we learn the parameters and structure of a graphical model?

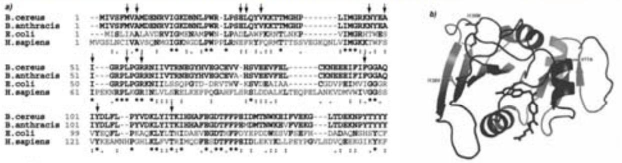


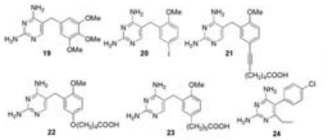
52 months
90 months
1. Find a target
2. Identify hits
3.Hit-to-lead characterize hits
4. Lead optimization and synthesis
5. Assay
$500,000,000 to $2,000,000,000

The chemical space

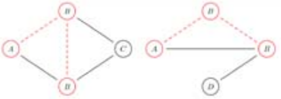
- 10^{60} possible small organic molecules
- 10^{22} stars in the observable universe

Example Question: Predicting Function from Structure





19 20 21 22 23 24



How similar are two graphs? How similar is their structure? How similar are their node and edge labels?

Joska, T. M. & Anderson, A. C. 2006. Structure-activity relationships of Bacillus cereus and Bacillus anthracis dihydrofolate reductase: toward the identification of new potent drug leads. Antimicrobial agents and chemotherapy, 50, 3435-3443.

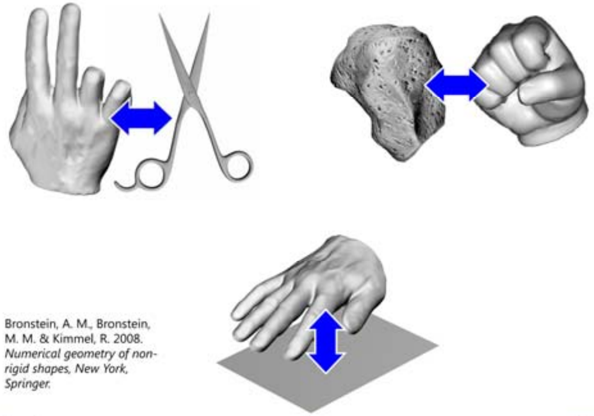
Graph Comparison

- Similar Property Principle: Molecules having similar structures should have similar activities.
- Structure-based representations: Compare molecules by comparing substructures, e.g.
  - Sets as vectors: Measure similarity by the cosine distance
  - Sets as sets: Measure similarity by the Jaccard distance
  - Sets as points: Measure similarity by Euclidean distance
- Problems: Dimensionality, Non-Euclidean cases

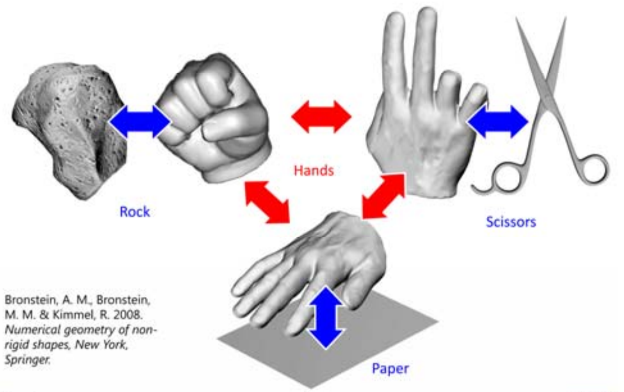
4) Little Excursus: What is similarity?

What is Similar?

Image credit to Eamonn Keogh (2008)



Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

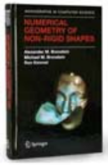


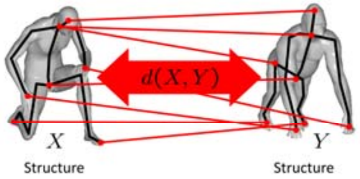
Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

Similarity and Correspondence

Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

http://www.inf.usi.ch/bronstein/

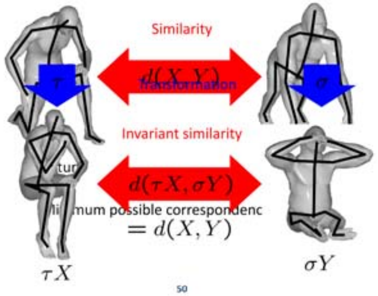




Correspondence quality = structure similarity (distortion)

Minimum possible correspondence distortion

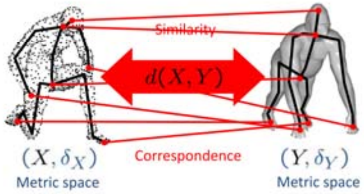
Invariant Similarity



Gromov-Hausdorff dist: finding the opt. correspondence

Michail Gromov (1943-)

Gromov, M. (1984) Infinite groups as geometric objects.



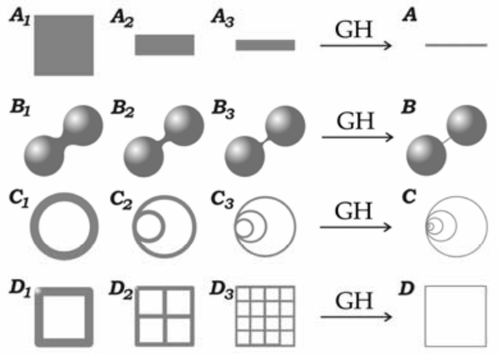
Felix Hausdorff (1868-1942)

$$d_{\mathrm{GH}}(X,Y) = \frac{1}{2} \min_{\mathcal{C}} \max_{\substack{(x_i,y_i)\in\mathcal{C} \\ (x_j,y_j)\in\mathcal{C}}} |\delta_X(x_i,x_j) - \delta_Y(y_i,y_j)|$$

$$\forall x_i \, \exists y_i \text{ s.t.} (x_i,y_i) \in \mathcal{C} \qquad \forall y_i \, \exists x_i \text{ s.t.} (x_i,y_i) \in \mathcal{C}$$

Discrete optimization over correspondences is NP hard !

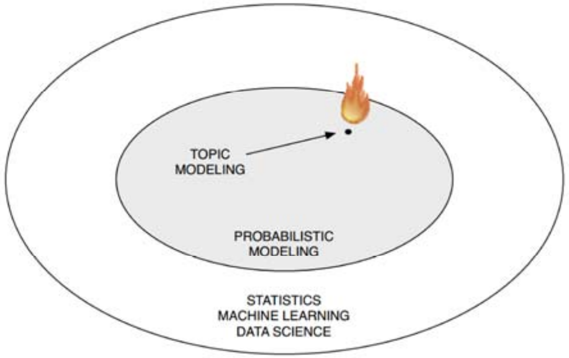
Example



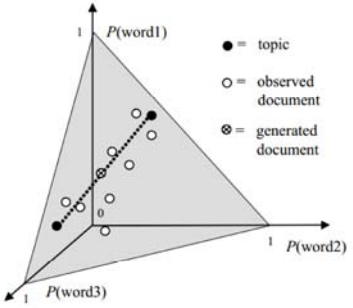
Sormani, C. 2010. How Riemannian Manifolds Converge: A Survey. arXiv preprint arXiv:1006.0411.

5) Probabilistic Topic Models

Topic modelling – small topic but hot topic in ML

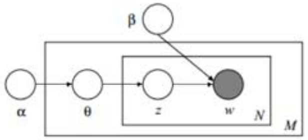


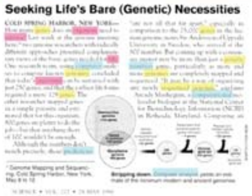
Geometry of Topic Models



Blei, D. M. 2012. Probabilistic topic models. Communications of the ACM, 55, (4), 77-84, doi:10.1145/2133806.2133826.

Generative statistical model for natural language



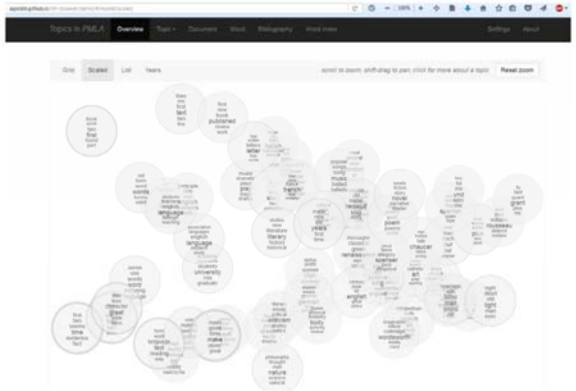

Seeking Life's Bare (Genetic) Necessities

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

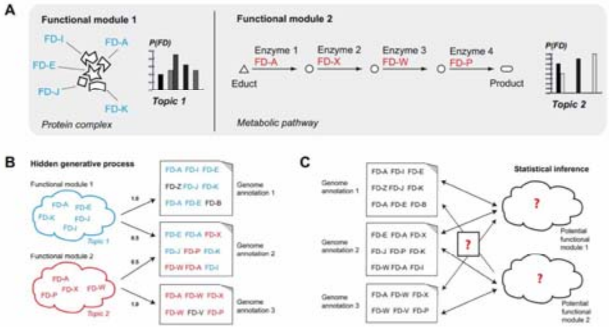
Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. The Journal of machine Learning research JMLR, 3, 993-1022.

Motivation: to get insight into unknown document sets



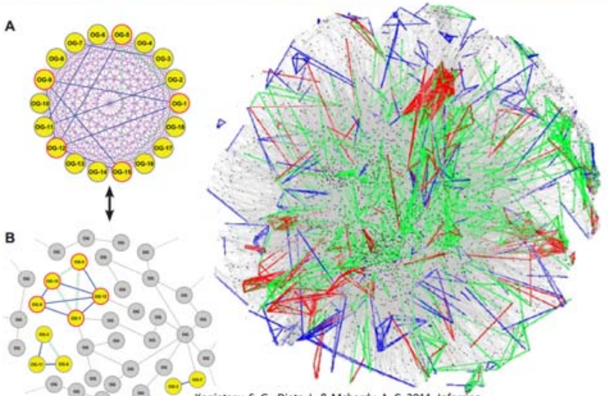
http://agoldst.github.io/dfr-browser/demo/#/model/scaled

Example from Bioinformatics



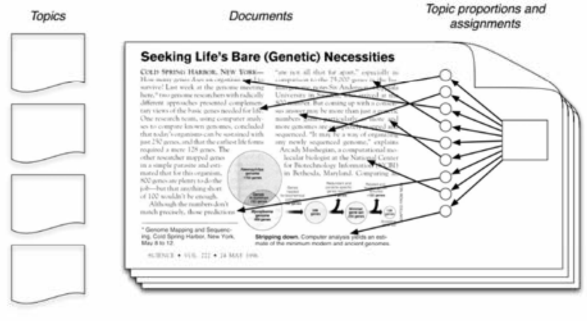
Konietzny, S. G., Dietz, L. & Mchardy, A. C. 2011. Inferring functional modules of protein families with probabilistic topic models. BMC bioinformatics, 12, (1), 1.

Eval. scheme for inferred potential functional modules



Konietzny, S. G., Dietz, L. & Mchardy, A. C. 2011. Inferring functional modules of protein families with probabilistic topic models. BMC bioinformatics, 12, (1), 1.

Example

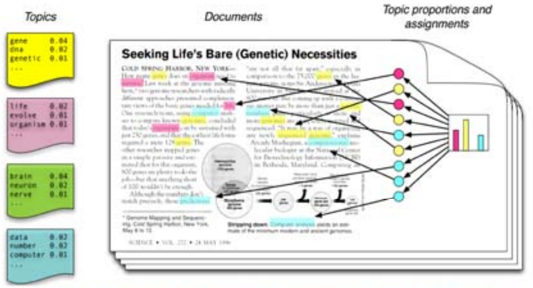

Topics
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities

We only observe the docs – the other structure is hidden; then we compute the posterior p(t,p,a|docs)

We find topics through a Generative Probabilistic Model

Goal: to get insight in unknown document collections

See a nice demo http://agoldst.github.io/dfr-browser/demo/#/model/grid

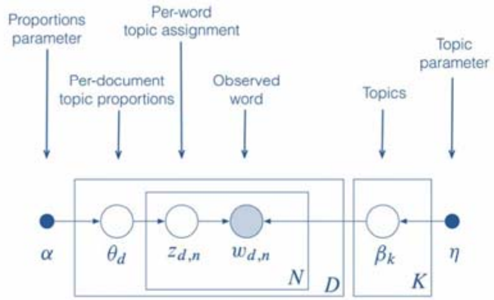


Each doc is a random mix of corpus-wide topics and each word is drawn from one of these topics

Output Example: 4 learned topics (details in Blei, 2008)

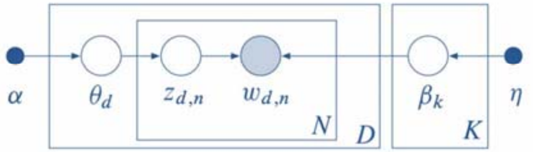
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA is an example for a probabilistic graphical model



- Encodes assumptions on data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)

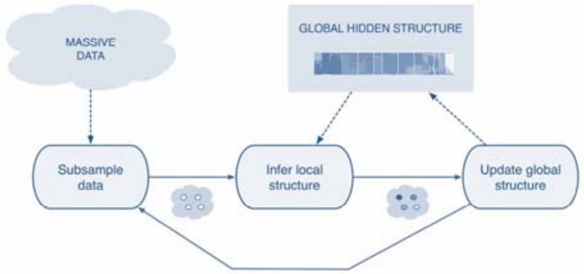
Posterior inference



$$p(\beta, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}$$

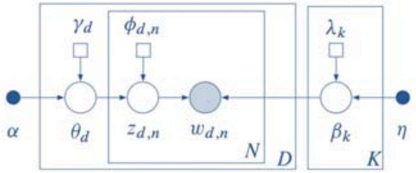
We can't compute the denominator, the marginal $p\ (w)$, therefore we use approximate inference; However, this do not scale well …

For "big data" stochastic variational inference



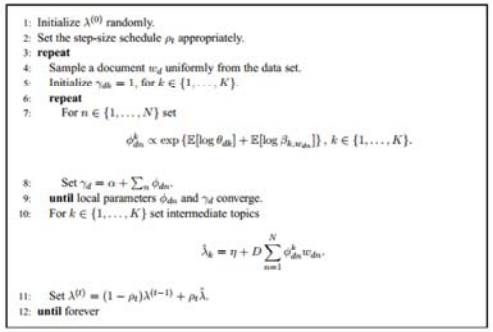
Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

Stochastic variational inference in LDA



1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form intermediate topics from those local parameters
4. Update topics as a weighted average of intermediate and current topics

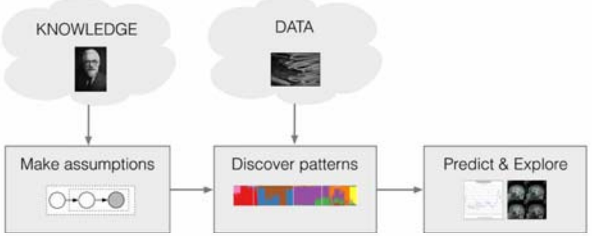
Stochastic variational inference



```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\rho_t$ appropriately.
repeat
  Sample a document $w_d$ uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
  repeat
    For $n \in \{1, \ldots, N\}$ set
        $\phi^k_{dn} \propto \exp\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\}, k \in \{1, \ldots, K\}$.
    Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters $\phi_{dn}$ and $\gamma_d$ converge.
  For $k \in \{1, \ldots, K\}$ set intermediate topics
        $\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi^k_{dn} w_{dn}$.
  Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever
```

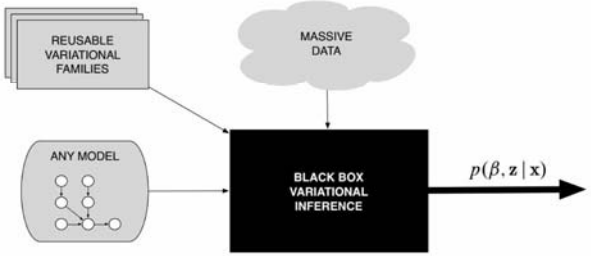
Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

Approximate inference can be difficult to achieve



- Approximate inference can be difficult to derive.
- Especially true for models that are not conditionally conjugate (Discrete choice models, Bayesian generalized linear models, ...)
- Holds us back from trying many models.

Black Box Approach


REUSABLE VARIATIONAL FAMILIES
MASSIVE DATA
ANY MODEL
BLACK BOX VARIATIONAL INFERENCE
$p(\beta, \mathbf{z} \mid \mathbf{x})$

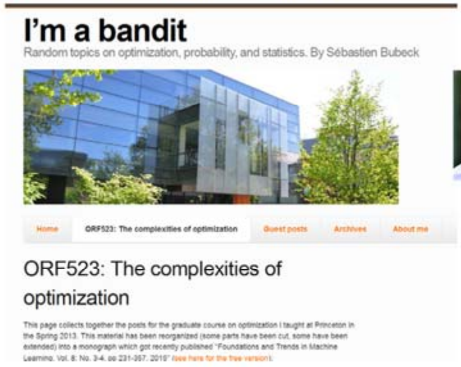
- Easily use variational inference with *any model*
- No exponential family requirements
- No mathematical work beyond specifying the model

Conclusion: What are future challenges

- **Flexible** and expressive components for building models are of utmost importance
- **Scalable** and generic inference algorithms (multi-task and transfer learning)
- **Usability** gets a totally new importance: Easy to use algorithms for the non-expert user to stretch probabilistic modeling into new areas
- Topic models are **one** approach towards detection of topics in document collections
- Example: Identifying re-occurring patterns in such data collections (gaining new knowledge)

6) Graph Bandits

The complexities of optimization: Sébasitien Bubeck



I'm a bandit

Random topics on optimization, probability, and statistics. By Sébastien Bubeck

ORF523: The complexities of optimization

This page collects together the posts for the graduate course on optimization I taught at Princeton in the Spring 2013. This material has been reorganized (some parts have been cut, some have been extended) into a monograph which got recently published "Foundations and Trends in Machine Learning, Vol. 8: No. 3-4, pp 231-357, 2015" (see here for the free version).

Bubeck, S. & Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Machine Learning, 5, (1), 1-122.

https://blogs.princeton.edu/imabandit/

Also very interesting: Bubeck, S. 2015. Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning, 8, (3-4), 231-357.

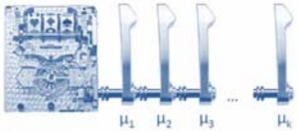
What is a bandit?

- Slot-machine (bandit - robs your money)
- One-armed bandit
- Very simple model for sequential decision making under uncertainty
- Main challenge: exploration versus exploitation
- Many application domains: A/B-Testing, Crowdsourcing, optimization, search, …

Multi-Armed Bandits problem

- Multi-armed bandit:= a gambler strategically operating multiple machines in order to draw the highest possible profits
- There are n slot-machines ("einarmige Banditen")
- Each machine i returns a reward $y \approx P(y;\, \Theta_i)$
- Challenge: The machine parameter Θ_i is unknown
- Which arm of a slot machine should a gambler pull to maximize his cumulative reward over a sequence of trials? (stochastic setting or adversarial setting)

Machine Parameters of the k-armed Bandit



Each arm a either

wins (reward=1) with fixed (unknown) probability μ_a, or

loses (reward=0) with fixed (unknown) probability $1 - \mu_a$

- All draws are independent given μ_1 … μ_k
- Problem:
  How to pull arms to maximize the total reward?

Underlying Principle of the k-Armed Bandits problem

- Let $a_t \in \{1, \ldots, n\}$ be the choice of a machine at time t
- Let $y_t \in \mathbb{R}$ be the outcome with a mean of $\langle y_{at} \rangle$
- Now, the given policy maps all history to a new choice:

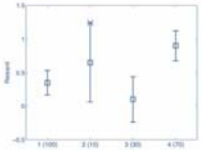
$$\pi :\ [(a_1, y_1), (a_2, y_2), \ldots, (a_{t\text{-}1}, y_{t\text{-}1})] \mapsto a_t$$

- The problem: Find a policy π that $\max\langle y_T \rangle$
- Now, two effects appear when choosing such machine:
  - You collect more data about the machine (=knowledge)
  - You collect reward
- Exploration and Exploitation
  - **Exploration:** Choose the next action a_t to $min\langle H(b_t)\rangle$
  - **Exploitation:** Choose the next action a_t to $max\langle y_t \rangle$
- models an agent that simultaneously attempts to acquire new knowledge (called "exploration") and optimize his or her decisions based on existing knowledge (called "exploitation"). The agent attempts to balance these competing tasks in order to maximize total value over the period of time considered.

More information: http://research.microsoft.com/en-us/projects/bandits

MAP-Principle: "Optimism in the face of uncertainty"

$$a_t = \max_{a \in \mathcal{A}} \left(\widehat{r}_t(a) + \sqrt{\frac{\log(1/\delta)}{T_t(a)}} \right)$$



$$a_t = \max_{a \in \mathcal{A}} \left(\mathrm{rew}_t(a) + \mathrm{uncert}_t(a) \right)$$

Exploitation
the higher the (estimated) reward the higher the chance to select the action

Exploration
the higher the (theoretical) uncertainty the higher the chance to select the action

Auer, P., Cesa-Bianchi, N. & Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47, (2-3), 235-256.

A bandit in a graph is still a bandit ☺

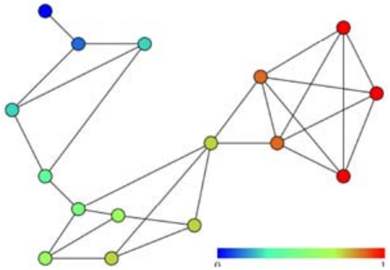
- Let G a known graph with K nodes $\{1, 2, \ldots, K\}$
- Let f be a unknown function defined on the set of nodes
- For $t = 1$ to n,
  - Select a node I_t
  - Observe reward $r_t = f(I_t) + \epsilon_t$
- Goal: maximize sum of expected rewards
- Equivalently minimize regret:

$$R_n = \sum_{t=1}^{n} (f^* - f(I_t)),$$

where $f^* = \max_{1 \leq i \leq K} f(i)$.

- We care about the case when $K > n$

Smooth Graph Function



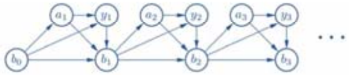
Knowledge Representation in MAB

- Knowledge can be represented in two ways:
- 1) as full history $\quad h_t = [(a_1, y_1), (a_2, y_2), \ldots, (a_{t\text{-}1}, y_{t\text{-}1})]$

  or
- 2) as belief $\quad b_t(\theta) = P(\theta | h_t)$

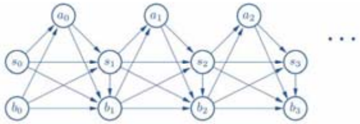
where Θ are the unknown parameters of all machines

The process can be modelled as belief MDP:



$$P(b'|y, a, b) = \begin{cases} 1 & \text{if } b' = b'_{[b,a,y]} \\ 0 & \text{otherwise} \end{cases}, \quad P(y|a, b) = \int_{\theta_a} b(\theta_a)\, P(y|\theta_a)$$

The optimal policies can be modelled as belief MDP



$$P(b'|s', s, a, b) = \begin{cases} 1 & \text{if } b' = b[s', s, a] \\ 0 & \text{otherwise} \end{cases}, \quad P(s'|s, a, b) = \int_{\theta} b(\theta)\, P(s'|s, a, \theta)$$

$$V(b, s) = \max_{a} \left[\mathsf{E}(r|s, a, b) + \textstyle\sum_{s'} P(s'|a, s, b)\, V(s', b') \right]$$

Poupart, P., Vlassis, N., Hoey, J. & Regan, K. An analytic solution to discrete Bayesian reinforcement learning. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 697-704.

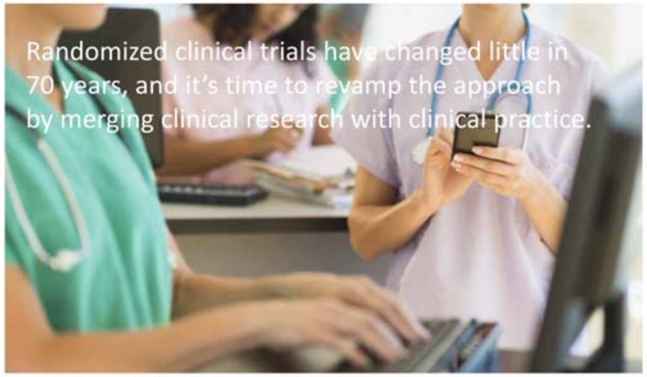
Applications

- Clinical trials: potential treatments for a disease to select from new patients or patient category at each round, see:

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Bulletin of the American Mathematics Society, vol. 25, pp. 285–294, 1933.

- Games: Different moves at each round, e.g. GO
- Adaptive routing: finding alternative paths, also finding alternative roads for driving from A to B
- Advertisement placements: selection of an ad to display at the Webpage out of a finite set which can vary over time, for each new Web page visitor

Multi-Armed Bandits and Clinical Trials



Randomized clinical trials have changed little in 70 years, and it's time to revamp the approach by merging clinical research with clinical practice.

http://fortune.com/2015/10/26/cancer-clinical-trial-belmont-report/

Example: Clinical Trials in rare diseases

Limitations of drug design for rare diseases due to:

- Lack of understanding of the underlying principles of the rare disease
  - Motivation: Research advances
- Unbalanced economic motivation (cost/benefit)
  - Motivation: Orphan Drug Act and other regulations
- Unavailability of # patients for standard trials
  - This is the true bottleneck!

Villar, S. S., Bowden, J. & Wason, J. 2015. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. 199-215, doi:10.1214/14-STS504.

Why should Bandit Strategies be of help here?

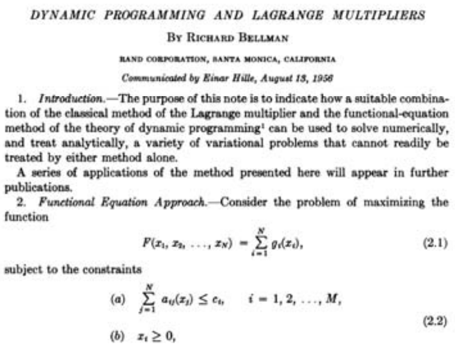
- The goal of Standard Randomized Controlled Trials (RCT) are a controlled learning setting:
  - Control for Type I and Type II errors, dependent of trial size n_{RCT}
  - In the case if the patient population N is smaller than the trial size n_{RCT}: underpowered trial – problem!
- If we change the goal to
- *"learning sufficient - to treat N as effectively as possible",*
- then **bandit strategies** – optimal policy for max. the expected reward - are perfectly suited!

Kuleshov, V. & Precup, D. 2014. Algorithms for multi-armed bandit problems. *arXiv:1402.6028*.

Learning vs. Earning dilemma

- Learning → experimenting with all treatments
- Earning → selecting one treatment only, based on experimentation results
- **Question 1: How much learning is best – for an optimal treatment of *N* patients?**
- Suppose N patients with a rare disease:
  - Experimental Group E and control group C
  - e.g. control = response rate pc and little information about experimental group
- **Question 2: How many allocations of treatment to E are necessary (= how much experimentation?)**

Answer to Q2: Dynamic Programming



DYNAMIC PROGRAMMING AND LAGRANGE MULTIPLIERS

By Richard Bellman

RAND CORPORATION, SANTA MONICA, CALIFORNIA

Communicated by Einar Hille, August 13, 1956

1. *Introduction.*—The purpose of this note is to indicate how a suitable combination of the classical method of the Lagrange multiplier and the functional-equation method of the theory of dynamic programming[1] can be used to solve numerically, and treat analytically, a variety of variational problems that cannot readily be treated by either method alone.

A series of applications of the method presented here will appear in further publications.

2. *Functional Equation Approach.*—Consider the problem of maximizing the function

$$F(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} g_i(x_i), \quad (2.1)$$

subject to the constraints

$$\begin{aligned} &(a) \quad \sum_{j=1}^{N} a_{ij}(x_j) \le c_i, \qquad i = 1, 2, \ldots, M, \\ &(b) \quad x_i \ge 0, \end{aligned} \quad (2.2)$$

Richard Ernest BELLMAN (1920-1984)

Bellman, R. 1956. Dynamic programming and Lagrange multipliers. Proceedings of the National Academy of Sciences, 42, (10), 767-769.

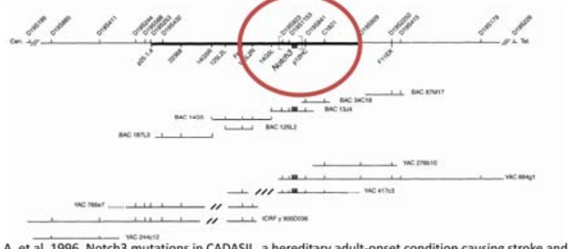
Some Statistics on rare diseases (orphan diseases)

- 7,000 + different types - more being discovered every day
- >10% of the world population is suffering (if all of the people with rare diseases lived in one country, it would be the world's 3rd most populous country)
- 80% of rare diseases are genetic, so are present throughout a person's lifetime, even if symptoms do not immediately appear
- >50% of the people affected by rare diseases are children
- Are responsible for 35% of deaths in the first year of life
- The prevalence distribution is skewed – 80% of all rare disease patients are affected by 350 rare diseases
- >50% of rare diseases do not have a disease specific foundation supporting or researching their rare disease

https://globalgenes.org/rare-diseases-facts-statistics/
https://www.hon.ch/HONselect/RareDiseases/

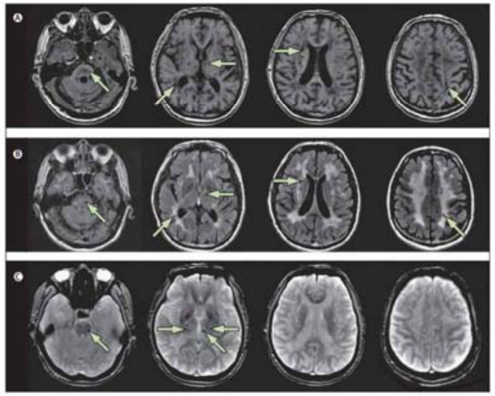
Example Rare Disease: CADASIL

- Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy
- is a hereditary disease affecting all the small cerebral arteries. It causes subcortical infarcts and damages the white matter (leukoencephalopathy) and it is due to various mutations of the Notch3 gene situated on chromosome 19:



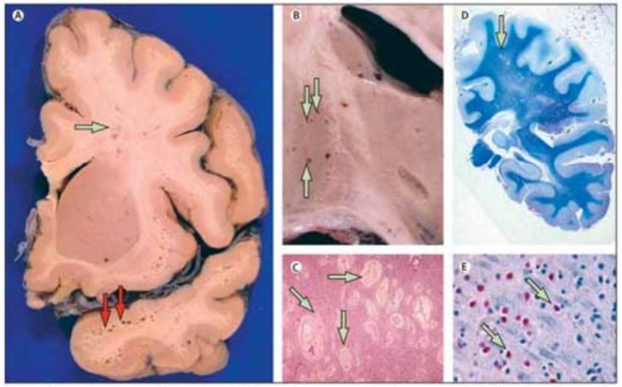
Joutel, A. et al. 1996. Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature, 383, (6602), 707-710, doi:10.1038/383707a0.

CADASIL



Chabriat, H., Joutel, A., Dichgans, M., Tournier-Lasserve, E. & Bousser, M.-G. 2009. CADASIL. The Lancet Neurology, 8, (7), 643-653, doi:http://dx.doi.org/10.1016/S1474-4422(09)70127-9.

CADASIL



Chabriat, H., Joutel, A., Dichgans, M., Tournier-Lasserve, E. & Bousser, M.-G. 2009. CADASIL. The Lancet Neurology, 8, (7), 643-653, doi:http://dx.doi.org/10.1016/S1474-4422(09)70127-9.

Conclusion and Future Challenges

Final conclusion

- Bandit strategy: Is experimentation worth it for a small number N?
- Reconcile clinical trials and clinical practice
- Extensions should deal with randomization, delayed responses and uncertainty around N
- Bayesian bandits need Online-ML
- Bandits are a great source of inspirations and building blocks for solving many problems
- Future work: convex optimization, contextual, combinatorial, …

Berry, D. A. & Fristedt, B. 1985. Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability), Springer.

Thank you!

Questions

Sample Questions (1/3)

- What kind of graphical models are used in medical informatics?
- Which type of graph is particularly useful for inference and learning?
- What is the key challenge in the application of graphical models for health informatics?
- What was Judea Pearl (1988) discussing in his paper, for which he received the Turing award?
- What main difficulties arise during breast cancer prognosis?
- What can be done to increase the robustness of prognostic cancer tests?
- Inference in Bayes Nets is NP-complete, but there are certain cases where it is tractable, which ones?

Sample Questions (2/3)

- Why do we want to apply ML to graphs?
- Describe typical ML tasks on the example of blood cancer cells!
- If you have a set of points – which similarity measures are useful?
- What is the advantage of factor graphs?
- Why is the Gromov-Hausdorff distance useful?
- What is the central goal of a generative probabilistic model?
- Describe the LDA-model and its application for topic modelling!

Sample Questions (2/3)

- Briefly describe the stochastic variational inference algorithms!
- What is the principle of a bandit?
- How does a multi-armed bandit (MAB) work?
- In which ways can a MAB represent knowledge?
- What is the main problem of a clinical trail – and maybe the main problem in clinical medicine?
- Why are rare diseases both important and relevant? Describe an example disease!
- What is the big problem in clinical trials for rare diseases?
- What did Richard Bellman (1956) describe with dynamic programming?
- Why are graph bandits a hot topic for ML research?

Solutions of the Quiz

- 1=this is a factor graph of an undirected graph – we have seen this in protein networks (refer to slide Nr. 70 in lecture 5). Factor graph is bipartite and has two types of nodes: Variables, which can be either evidence variables (when we know its value) or query variables (when the value is unknown and we want to predict the value); and factors, which define the relationship between variables in the graph. Each factor can be connected to many variables and comes with a factor function to define the relationship between these variables. For example, if a factor node is connected to two variables nodes A and B, a possible factor function could be imply(A,B), meaning that if the random variable A takes value 1, then so must the random variable B. Each factor function has a weight associated with it, which describes how much influence the factor has on its variables in relative terms. For more information please consult: http://deepdive.stanford.edu/inference
- 2= this is the decomposition of a tree, rooted at nodes into subtrees
- 3= metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions. With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human. Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG)[1], EcoCyc [2], BioCyc [3] and metaTIGER [4]. Metabolic networks are powerful tools for studying and modelling metabolism.
- 4= MYCIN –expert system that used early AI (rule-based) to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight — the name derived from the antibiotics themselves, as many antibiotics have the suffix "-mycin".
- 5= Protein-Protein Interaction network (undirected graph here)
- 6= PPI with critical node, bottleneck, hub, etc.

Appendix

GM are amongst the most important ML developments

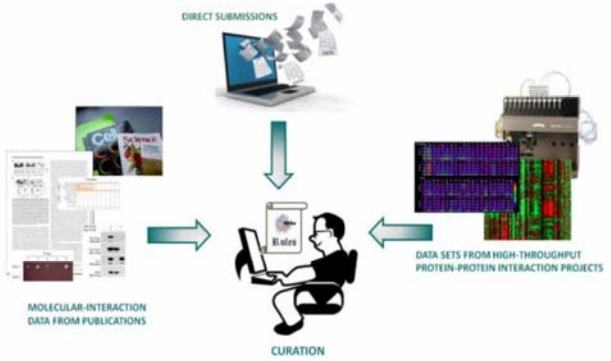
- Key Idea: Conditional independence assumptions are very useful – however: Naïve Bayes is extreme!
- X is *conditionally independent* of Y, given Z, if the P(X) governing X is independent of value Y, given value of Z:

$$(\forall i,j,k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

can be abbr. with $P(X|Y,Z) = P(X|Z)$

- Graphical models express sets of conditional independence assumptions via graph structure
- The graph structure plus associated parameters define joint probability distribution over the set of variables

Where do the data come from?

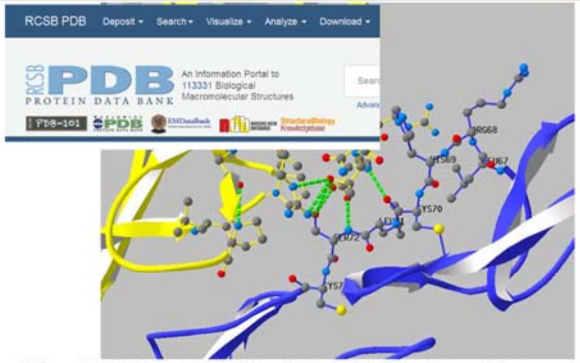

DIRECT SUBMISSIONS
DATA SETS FROM HIGH-THROUGHPUT PROTEIN-PROTEIN INTERACTION PROJECTS
MOLECULAR-INTERACTION DATA FROM PUBLICATIONS
CURATION

http://www.ebi.ac.uk/intact/

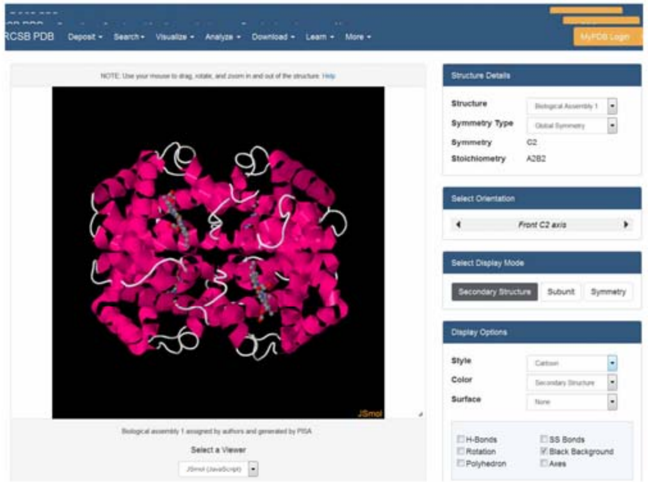
Slide 4-19: Biomedical databases …

- … are libraries of life science data, collected from scientific experiments and computational analyses.
- … contain (clinical, biological, …) data from clinical work, genomics, proteomics, metabolomics, microarray gene expression, phylogenetics, etc.
- Examples:
  - Text: e.g. PubMed, OMIM (Online Mendelian Inheritance in Man);
  - Sequence data: e.g. Entrez, GenBank (DNA), UniProt (protein).
  - Protein structures: e.g. PDB, Structural Classification of Proteins (SCOP), CATH (Protein Structure Classification);

Slide 4-20 Example Database: PDB



Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*



NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure. Help

Structure Details

Structure: Biological Assembly 1

Symmetry Type: Global Symmetry

Symmetry: C2

Stoichiometry: A2B2

Select Orientation: Front C2 axis

Select Display Mode: Secondary Structure

Display Options

Style: Cartoon

Color: Secondary Structure

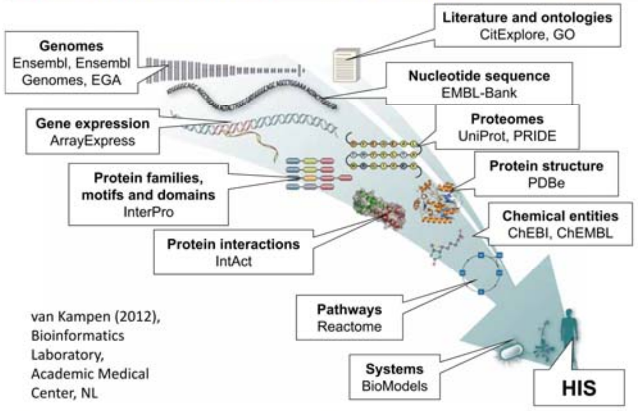
Surface: None

□ H-Bonds □ SS Bonds
□ Rotation ☑ Black Background
□ Polyhedron □ Axes

Biological assembly 1 assigned by authors and generated by PISA

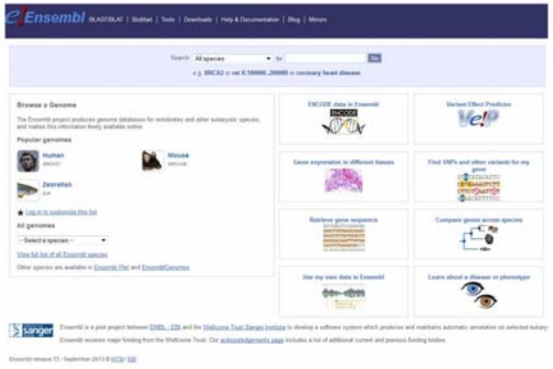
Select a Viewer: JSmol (JavaScript)

Slide 4-21 Databases: From Molecules to Systems



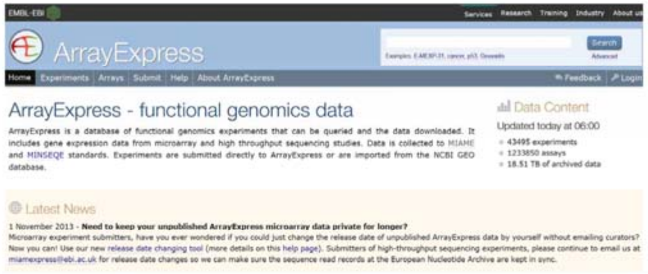
van Kampen (2012), Bioinformatics Laboratory, Academic Medical Center, NL

Slide 4-22: Example Genome Database: Ensembl


Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search: All species
Go
Browse a Genome
Popular genomes
Human
Mouse
Zebrafish
All genomes
- Select a species -
ENCODE data in Ensembl
Variant Effect Predictor
VeP
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
Learn about a disease or phenotype
sanger

http://www.ensembl.org/index.html

Slide 4-23 Ex. Gene Expression Database: ArrayExpress



ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content

Updated today at 06:00

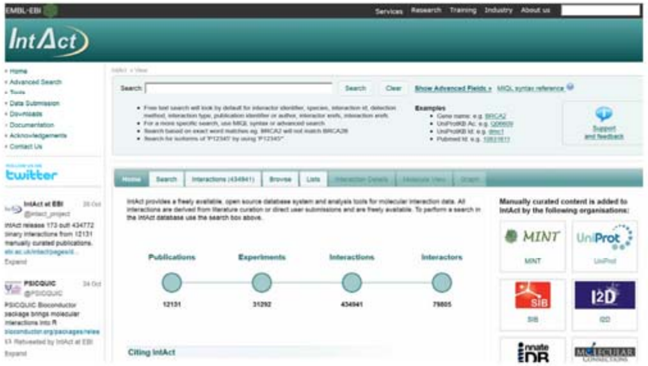
- 43495 experiments
- 1233850 assays
- 18.51 TB of archived data

Latest News

1 November 2013 - **Need to keep your unpublished ArrayExpress microarray data private for longer?**
Microarray experiment submitters, have you ever wondered if you could just change the release date of unpublished ArrayExpress data by yourself without emailing curators? Now you can! Use our new release date changing tool (more details on this help page). Submitters of high-throughput sequencing experiments, please continue to email us at miamexpress@ebi.ac.uk for release date changes so we can make sure the sequence read records at the European Nucleotide Archive are kept in sync.

http://www.ebi.ac.uk/arrayexpress/

Slide 4-24: Example Protein Interaction Database: IntAct



EMBL-EBI

Services Research Training Industry About us

IntAct

- Home
- Advanced Search
- Tools
- Data Submission
- Downloads
- Documentation
- Acknowledgements
- Contact Us

twitter

IntAct at EBI 28 Oct @intact_project
IntAct release 173 out! 434772 binary interactions from 12131 manually curated publications. ebi.ac.uk/intact/pages/d... Expand

PSICQUIC 24 Oct @PSICQUIC
PSICQUIC Bioconductor package brings molecular interactions into R bioconductor.org/packages/relea... Retweeted by IntAct at EBI Expand

Search: Search Clear Show Advanced Fields » MIQL syntax reference

- Free text search will look by default for interactor identifier, species, interaction id, detection method, interaction type, publication identifier or author, interactor xrefs, interaction xrefs.
- For a more specific search, use MIQL syntax or advanced search
- Search based on exact word matches eg. BRCA2 will not match BRCA2B
- Search for isoforms of 'P12345' by using 'P12345*'

Examples
- Gene name: e.g. BRCA2
- UniProtKB Ac: e.g. Q06609
- UniProtKB Id: e.g. dmc1
- Pubmed Id: e.g. 10831611

Support and feedback

Home | Search | Interactions (434941) | Browse | Lists | Interaction Details | Molecule View | Graph

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. To perform a search in the IntAct database use the search box above.

Publications	Experiments	Interactions	Interactors
12131	31292	434941	79805

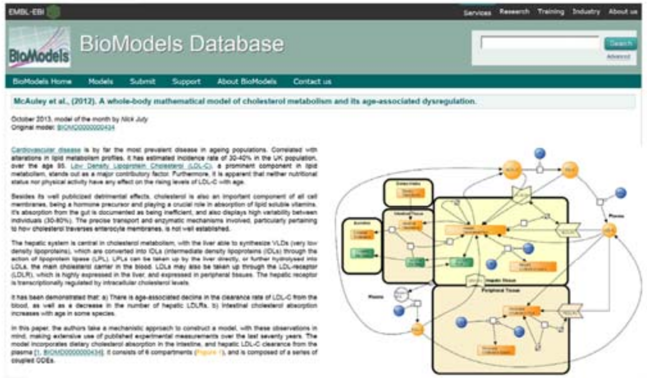
Citing IntAct

Manually curated content is added to IntAct by the following organisations:

MINT UniProt SIB I2D

http://www.ebi.ac.uk/intact/

Slide 4-25: Example for Systems Database: BioModels



EMBL-EBI

Services Research Training Industry About us

BioModels Database

Search Advanced

BioModels Home Models Submit Support About BioModels Contact us

McAuley et al., (2012). A whole-body mathematical model of cholesterol metabolism and its age-associated dysregulation.

October 2013, model of the month by Nick Juty
Original model: BIOMD0000000434

Cardiovascular disease is by far the most prevalent disease in ageing populations. Correlated with alterations in lipid metabolism profiles, it has estimated incidence rate of 30-40% in the UK population, over the age 55. Low Density Lipoprotein Cholesterol (LDL-C), a prominent component in lipid metabolism, stands out as a major contributory factor. Furthermore, it is apparent that neither nutritional status nor physical activity have any effect on the rising levels of LDL-C with age.

Besides its well publicized detrimental effects, cholesterol is also an important component of all cell membranes, being a hormone precursor and playing a crucial role in absorption of lipid soluble vitamins. It's absorption from the gut is documented as being inefficient, and also displays high variability between individuals (30-80%). The precise transport and enzymatic mechanisms involved, particularly pertaining to how cholesterol traverses enterocyte membranes, is not well established.

The hepatic system is central in cholesterol metabolism, with the liver able to synthesize VLDs (very low density lipoproteins), which are converted into IDLs (intermediate density lipoproteins (IDLs) through the action of lipoprotein lipase (LPL). LPLs can be taken up by the liver directly, or further hydrolysed into LDLs, the main cholesterol carrier in the blood. LDLs may also be taken up through the LDL-receptor (LDLR), which is highly expressed in the liver, and expressed in peripheral tissues. The hepatic receptor is transcriptionally regulated by intracellular cholesterol levels.

It has been demonstrated that: a) There is age-associated decline in the clearance rate of LDL-C from the blood, as well as a decrease in the number of hepatic LDLRs. b) Intestinal cholesterol absorption increases with age in some species.

In this paper, the authors take a mechanistic approach to construct a model, with these observations in mind, making extensive use of published experimental measurements over the last seventy years. The model incorporates dietary cholesterol absorption in the intestine, and hepatic LDL-C clearance from the plasma [1, BIOMD0000000434]. It consists of 6 compartments (Figure 1), and is composed of a series of coupled ODEs.

http://www.ebi.ac.uk/biomodels-main/

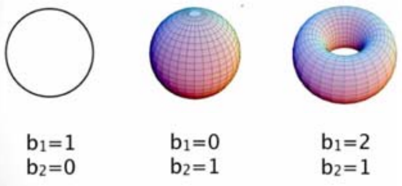
Distinguish topological spaces

Enrico Betti (1823-1892)

Counts the number of "i-dimensional holes"

bi is the "i-th Betti number"

Emmy Noether (1882-1935)



$b_1=1$	$b_1=0$	$b_1=2$
$b_2=0$	$b_2=1$	$b_2=1$

Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)

Zomorodian, A. & Carlsson, G. 2005. Computing Persistent Homology. *Discrete & Computational Geometry, 33, (2), 249-274.*

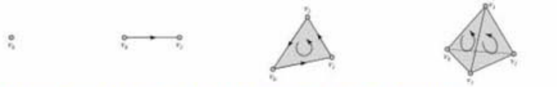
Structural Patterns are often hidden in weakly str. data

- Statement of Vin de Silva (2003), Pomona College:
- Let M be a topological or metric space, known as the *hidden parameter space*;
- let $\mathbb{R}^d$ be a Euclidean space, the *observation space*,
- and let $f: M \longrightarrow \mathbb{R}^d$ be a continuous embedding.
- Furthermore, let $X \subset M$ be a finite set of data points, perhaps the realization of a stochastic process, i.e., a family of random variables $\{X_i, i \in I\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, and denote $Y = f(X) \subset \mathbb{R}^d$ the images of these points under the mapping f.
- We refer to X as *hidden data*, and Y as the *observed data*.
- M, f and X are unknown, but Y is - so can we identify M?

De Silva, V. 2004. GEOMETRY AND TOPOLOGY OF POINT-CLOUD DATA SETS: A STATEMENT OF MY RESEARCH INTERESTS.

https://www.pomona.edu/directory/people/vin-de-silva

Topological Data Mining



- Mega Problem: To date none of our known methods, algorithms and tools scale to the massive amount and dimensionalities of data we are confronted in practice;
- we need much more research efforts towards making computational topology successful as a general method for data mining and knowledge discovery

Holzinger, A. 2014. On Topological Data Mining. In: Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 331-356, doi:10.1007/978-3-662-43968-5_19.

Topic model toolkits

- Particular topic models
  - Stanford topic model toolbox http://nlp.stanford.edu/software/tmt
  - Topic modeling at Princeton http://www.cs.princeton.edu/~blei/topicmodeling.html
  - MALLET (Java) http://mallet.cs.umass.edu
  - Network topic models: Bayes-stack https://github.com/bgamari/bayes-stack
  - Gensim (Python) http://radimrehurek.com/gensim/
  - R package for Topic models. http://epub.wu.ac.at/3987/
- Frameworks for generative models
  - Variational inference: Infer.net http://research.microsoft.com/infernet/
  - Gibbs sampling: OpenBUGS http://openbugs.net/