computational scientific discovery

ljupčo todorovski
jožef stefan institute, department of knowledge technologies
ljubljana, slovenia

tutorial on the discovery science conference 2005
tutorial outline: from specific to general

- FIRST PART: computational discovery of equation-based models
  - brief history
  - explanatory vs. predictive models
  - examples of applications
  - recent work and open issues

- SECOND PART: other aspects
  - inference of genetic/chemical reaction networks
  - literature-based discovery

- CONCLUSION: what is computational scientific discovery?
FIRST PART
computational discovery of equation-based models
motivation: automated modeling

- computational support for building equation-based models

- existing modeling (system identification) methods:
  - assume that the model structure is known
  - or assume linear/NN structure (if the model is unknown)
  - task: identify the values of the model parameters

- in contrast, computational discovery methods:
  - does not assume (a single) prescribed model structure
  - task: identify both structure and parameters of the model
motivation: inducing explanatory models

- two approaches to modeling of real-world systems:
  1. theoretical (knowledge-driven): expert derives a structure of the model based on domain-specific knowledge (first principles)
     result: explanatory white-box models
  2. empirical (data-driven): try different models to fit observations
     result: predictive black-box models

- learning/mining methods usually support 2.

- computational discovery methods integrate 1. and 2.:
  - induction from observations
  - result: predictive AND explanatory models
task definition and brief history of the field
equation discovery: task definition

- GIVEN:
  - a set of system variables \( V = \{v_1, v_2, \ldots, v_n\} \)
  - a table of measurements (observations) of the system variables

\[
\begin{array}{cccc}
  v_1 & v_2 & \cdots & v_n \\
  v_{1,1} & v_{2,1} & \cdots & v_{n,1} \\
  v_{1,2} & v_{2,2} & \cdots & v_{n,2} \\
  \vdots & \vdots & \cdots & \vdots \\
  v_{1,m} & v_{2,m} & \cdots & v_{n,m} \\
\end{array}
\]

- FIND equations that summarize the observations:

\[ F(v_1, v_2, \ldots, v_n) = 0 \]
example: Kepler’s law (1)

• given experimental/observational data
  – planet’s orbital period vs. its distance from the Sun
example: Kepler’s law (2)

- (re-)discover Kepler’s Third Law

\[ \frac{d^3}{p^2} = \text{const} \]
example: population dynamics (1)

- given time-series measurement data
  - concentration of species through time
example: population dynamics (2)

- find model Lotka-Volterra model (ordinary differential equations)

\[ \frac{d}{dt} \text{hare} = 2.5 \cdot \text{hare} - 0.3 \cdot \text{hare} \cdot \text{fox} \]

\[ \frac{d}{dt} \text{fox} = 0.1 \cdot 0.3 \cdot \text{hare} \cdot \text{fox} - 1.2 \cdot \text{fox} \]
equation discovery methods: brief historical overview (1)

• relatively simple data-driven heuristics
  – fahrenheit/ef [Żytkow, 1987; Zembowitz & Żytkow 1992]
  – ids [Nordhausen & Langley 1990]

• measurement units and scale types
  – coper [Kokar, 1985]
  – sds [Washio & Motoda, 1997]

• abacus family (heuristic search, piecewise models)
  – abacus [Falkenheimer & Michalski, 1986]
  – arc [Moulet, 1992]

• other/misc: e* [Schaffer, 1990]
bacon’s data-driven heuristics: Kepler’s Law example

- three simple heuristics:
  - detect variables with a constant value
  - if the increase of $X$ leads to increase of $Y$ introduce $X/Y$
  - if the increase of $X$ leads to decrease of $Y$ introduce $X \cdot Y$

<table>
<thead>
<tr>
<th>planet</th>
<th>$p$</th>
<th>$d$</th>
<th>$d/p = d/p$</th>
<th>$d^2/p = d \cdot (d/p)$</th>
<th>$d^3/p^2 = (d^2/p) \cdot (d/p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>0.39</td>
<td>0.24</td>
<td>1.62</td>
<td>0.63</td>
<td>1.03</td>
</tr>
<tr>
<td>Venus</td>
<td>0.72</td>
<td>0.62</td>
<td>1.16</td>
<td>0.84</td>
<td>0.97</td>
</tr>
<tr>
<td>Earth</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Mars</td>
<td>1.52</td>
<td>1.88</td>
<td>0.81</td>
<td>1.23</td>
<td>0.99</td>
</tr>
<tr>
<td>Jupiter</td>
<td>5.20</td>
<td>11.9</td>
<td>0.44</td>
<td>2.27</td>
<td>0.99</td>
</tr>
<tr>
<td>Saturn</td>
<td>9.54</td>
<td>29.5</td>
<td>0.32</td>
<td>3.08</td>
<td>1.00</td>
</tr>
</tbody>
</table>
equation discovery systems: typical laws/models (re-)discovered

- Kepler’s law (bacon): \( \frac{d^3}{p^2} = k \)

- ideal gas law (bacon): \( PV = kNT \)

- Black’s law (bacon, fahrenheit, ids):
  \[ c_1m_1T_1 + c_2m_2T_2 = T(c_1m_1 + c_2m_2) \]

- conservation of momentum (abacus):
  \[ m_1v_{11} + m_2v_{21} = m_1v_{12} + m_2v_{22} \]

- electrical circuit model (sds):
  \[
  \left( \frac{R_3h_{fe_2}}{R_3h_{fe_2}+h_{ie_2}} \frac{R_2h_{fe_1}}{R_2h_{fe_1}+h_{ie_1}} \frac{rL^2}{rL^2+R_1} \right) V - \frac{Q}{C} - \frac{Kh_{ie_3}X}{Bh_{fe_3}} = 0
  \]
equation discovery methods: brief historical overview (2)

- lagrange [Džeroski & Todorovski, 1993]
  differential equations, multivariate polynomials

- lagramge [Todorovski & Džeroski, 1997-2003]
  user specified space of candidate models, explanatory models

- pret [Bradley et al., 2001]
  compositional approach to modeling

- ipm/hipm [Langley et al., 2002; Todorovski et al., 2005]
  process-based (explanatory) models

- scaletrack [Washio et al., 2005]
  measurement units, dealing with unknown dimension of the system
rest of FIRST PART

- inducing explanatory models based on domain-specific knowledge
  - lagramge — using grammars to constrain the space of models
  - ipm/hipm — inducing process-based models

- other approaches
  - sds — knowledge about measurement scale types
  - pret — compositional modeling

- recent advances and open issues
inducing explanatory models
explanatory aspect of models: population dynamics

- find model Lotka-Volterra model (ordinary differential equations)
  \[
  \frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox \\
  \frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox
  \]

- three processes
  - exponential growth of hare population
  - exponential loss of fox population
  - predator-prey interaction between two species
how to induce explanatory models from data?

- integrate knowledge about building models in the induction process
- two different approaches
  - transform knowledge into grammars (lagramge)
  - use knowledge about processes directly (ipm/hipm)
transforming knowledge into grammars: lagramge
lagramge: a declarative-bias approach to equation discovery

- space of candidate equations (language bias) is:
  - defined by the user (declarative)
  - based on domain-specific knowledge

- declarative bias formalism (how to specify space of equations):
  - context-free grammar $G$
  - used to specify structure of expression on the right-hand side
  - induced equations are of the form

$$\frac{d}{dt} v = E_G$$

where $E_G$ is an expression derived using $G$
grammars for common language biases

- universal grammar (arbitrary arithmetical expression):
  \[ E \rightarrow E + F \mid E - F \]
  \[ F \rightarrow F \cdot T \mid F/T \]
  \[ T \rightarrow (E) \mid V \mid const \]
  follow the relative priority of arithmetic operators

- multivariate polynomials:
  \[ P \rightarrow P + const \cdot T \mid const \cdot T \]
  \[ T \rightarrow T \cdot V \mid V \]
from grammars to expressions: parse trees

- polynomial expression \( \text{const} \cdot x \cdot y + \text{const} \cdot z \)
lagramge: heuristic search through the space of parse trees

- search space
  - parse trees (with limited depth)
  - each represent a single equation structure

- beam search strategy
  1. start with simplest parse tree
  2. evaluate parse trees (parameter estimation)
  3. place evaluated trees in the beam
  4. if the beam has not changed, stop and report the best model(s)
  5. expand/refine each parse tree on the beam
  6. goto second step
lagramge: parameter estimation

- for a given equation structure
  - find values of the constant parameters \((\textit{const})\) that
  - minimize the discrepancy between measurements and simulation

- gradient descent search
  - replace differentiation with integration (teacher forcing):
    \[
    \dot{V} = F(V) \rightarrow V(t_i) = \int_{t_0}^{t_i} F(V) dt
    \]

- two evaluation functions:
  - sum of squared errors (SSE)
  - MDL = SSE + expression complexity penalty
From knowledge to grammar: population dynamics (1)

- Population dynamics models based on:
  - Conservation of mass/energy principle
  - Influences of interactions/processes summed up

- Change of primary producer (PP) population modeled as

\[
\frac{dPP}{dt} = growth(PP) - grazing(SP, PP) - loss(PP)
\]

- PP — primary producer species (plant, e.g., phytoplankton)
- SP — secondary producer species (animal, e.g., zooplankton)

- Three processes: growth, grazing, and mortality
from knowledge to grammar: population dynamics (2)

double monod(double v, double c) return v / (v + c);

\[
\begin{align*}
PPChange & \rightarrow PPGrowth - Grazing - PPLoss \\
PPGrowth & \rightarrow const \cdot V_{PP} | const \cdot Limitation \cdot V_{PP} \\
Limitation & \rightarrow V_{Nutrient} | \text{monod}(V_{Nutrient}, \text{const}) \\
Grazing & \rightarrow const \cdot V_{PP} \cdot V_{SP} \\
PPLoss & \rightarrow const \cdot V_{PP}
\end{align*}
\]

\[
\begin{align*}
V_{PP} & \rightarrow \text{phytoplankton} \\
V_{SP} & \rightarrow \text{zooplankton} \\
V_{Nutrient} & \rightarrow \text{phosphorus} \mid \text{nitrogen}
\end{align*}
\]
experiments with synthetic data: the methodology

- reconstruction of known models:
  - ten simulation traces generated from random initial states
  - Gaussian noise added different levels: 1%, 2%, 5%, 10%, and 20%

- three criteria for evaluating the performance:
  1. number of candidate models considered
  2. matching the original model structure
  3. simulation error (rmse) of the induced models (cross-validated)
experiments with synthetic data: a summary of the results

- number of candidate models drastically reduced
  - due to the use of knowledge-based constraints
  - reduction factor from 16 to 500,000

- improved noise-robustness and convergence rate
  - due to the use of knowledge-based constraints
  - AND the use of bounds on constant parameters

- avoid inappropriate structure or parameters
  - inappropriate structure can lead to more accurate
  - but UNSTABLE models
experiments with synthetic data: convergence rate

successfull reconstruction rate (in %)

noise level

L2
L2-cdc
L2-cpb
L1

no-noise
1%
2%
5%
10%
20%

computational scientific discovery (30)
lagoon of Venice: task description

• environment and modeling task:
  – excessive inflow of nutrients (from waste water)
    7 miokg/year of nitrogen and 1.4 miokg/year of phosphorus
  – generates excessive growth of algae (mainly Ulva Rigida)
  – what are the limiting factors for algae growth?

• six variables measured on two locations:
  – concentrations of nutrients
    nitrogen in ammonia and nitrate, phosphorus in orthophosphate
  – concentration of dissolved oxygen
  – temperature
  – algal biomass
lagoon of Venice: results

• relatively high model errors (rmse – 86.2841; 157.537)
  – due to high measurement errors (order 20% – 50%)
  – predict most of the biomass concentration peaks and crashes

• comprehensible white-box models induced:
  – ecology expert can easily understand them
  – they reveal the limiting factors for algae growth
  – dissolved oxygen, nitrogen-based nutrients, and temperature

• comparison with other approaches:
  – black-box models induced with comparable predictive accuracy
  – from preprocessed data (noise filtering, feature construction)
lagoon of Venice: simulation of the induced model (1)
lagoon of Venice: simulation of the induced model (2)
Ringkøbing fjord: model completion task

- expert specified only part of the model structure:

\[ \dot{h} = \frac{f(a)}{A} (h_{sea} - h + h_0) + \frac{Q_f}{A} + g(W_{vel}, W_{dir}) \]

- two parts of the structure left unspecified:
  - gate opening \( f \) that depends on number of open gate parts
  - wind forcing \( g \) that depends on wind direction and speed

- experiments:
  - \( f \) – constant or polynomial
  - \( g \) – constant, polynomial, and trigonometric functions
  - maximal polynomial degree of 5
**Ringkøbing fjord: grammars**

- **partial model structure grammar:**
  
  \[
  \text{WaterLevelChange} \rightarrow (F/A) \cdot \text{SaltWaterDrive} + \text{FreshWaterFlow} + G \\
  \text{SaltWaterDrive} \rightarrow (h_{sea} - h + \text{const}) \\
  \text{FreshWaterFlow} \rightarrow Q_f/A
  \]

- **alternative completion grammars:**

  \begin{array}{c|c}
  \text{F0} & F \rightarrow \text{const} \\
  \text{F1} & F \rightarrow F + \text{const} \cdot T_F \mid \text{const} \cdot T_F \\
  & T_F \rightarrow T_F \cdot V_F \mid V_F \\
  & V_F \rightarrow a \\
  \text{G0} & G \rightarrow \text{const} \\
  \text{G1} & G \rightarrow G + \text{const} \cdot T_G \mid \text{const} \cdot T_G \\
  & T_G \rightarrow T_G \cdot V_G \mid V_G \\
  & V_G \rightarrow W_{vel} \mid W_{dir} \\
  \text{G2} & V_G \rightarrow W_{vel} \mid \sin(W_{dir}) \mid \cos(W_{dir})
  \end{array}
Ringkøbing fjord: results

<table>
<thead>
<tr>
<th>task</th>
<th>training RMSE</th>
<th>cross-validated RMSE</th>
<th>#CMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 + G0</td>
<td>0.0848</td>
<td>0.106</td>
<td>1</td>
</tr>
<tr>
<td>F1 + G1</td>
<td>0.0655</td>
<td>0.0931</td>
<td>378</td>
</tr>
<tr>
<td>F1 + G2</td>
<td>0.0585</td>
<td>0.0903</td>
<td>2184</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.0556</td>
<td>2.389</td>
<td>2801</td>
</tr>
</tbody>
</table>

- best model captures the long-term dynamics of the water level
- can also predict short-term changes (one hour or day)
- allows for comparison of wind and gate opening influence
- black-box polynomials overfit the training data
Ringkøbing fjord model: long-term dynamics

![Graph showing water level over time](image)

- Water level [m]
- Time [days]

Measured data
Model
Ringkøbing fjord model: short-term prediction

<table>
<thead>
<tr>
<th>prediction/simulation period</th>
<th>RMSE</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>one hour</td>
<td>0.0168</td>
<td>0.976</td>
</tr>
<tr>
<td>one day</td>
<td>0.0425</td>
<td>0.845</td>
</tr>
<tr>
<td>whole observation period</td>
<td>0.0585</td>
<td>0.659</td>
</tr>
</tbody>
</table>
CASA: model revision task

• CASA model:
  – global production of biogenic trace gases in the atmosphere
  – complex system of difference equations

• we focused on NPPc part of CASA (CASA-NPPc):
  – predicts monthly net production of carbon at a given location
  – system of algebraic equations

• data set used for revision consists of:
  – 8 variables
  – measured at 303 locations
**CASA: the initial model**

\[
NPP_c = \max(0, E \cdot IPAR) \\
E = 0.389 \cdot T1 \cdot T2 \cdot W \\
\quad T1 = 0.8 + 0.02 \cdot topt - 0.0005 \cdot topt^2 \\
\quad T2 = 1.1814/((1 + \exp(0.2 \cdot (TDIFF - 10))) \cdot (1 + \exp(0.3 \cdot (-TDIFF - 10)))) \\
\quad TDIFF = topt - tempc \\
W = 0.5 + 0.5 \cdot eet/PET \\
PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m \\
\quad A = 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239 \\
IPAR = FPAR\_FAS \cdot \text{monthly}\_solar \cdot SOL\_CONV \cdot 0.5 \\
\quad FPAR\_FAS = \min((SR\_FAS - 1.08)/srdiff, 0.95) \\
\quad SR\_FAS = (1 + fas\_ndvi/1000)/(1 - fas\_ndvi/1000) \\
\quad SOL\_CONV = 0.0864 \cdot \text{days}\_per\_month
\]
CASA: dependencies between variables

![Diagram with variables and their dependencies]
CASA: proposed alternatives

- experts identified four "weak" parts of the model:
  - equations for $E, T_1, T_2,$ and $SR_{FAS}$

- two alternatives for $E = 0.0389 \cdot T_1 \cdot T_2 \cdot W$:
  1. $E = \text{const} \cdot T_1 \cdot T_2 \cdot W$
  2. $E = \text{const} \cdot T_1^{\text{const}} \cdot T_2^{\text{const}} \cdot W^{\text{const}}$

- two alternatives for $T_1 = 0.8 + 0.02 \cdot \text{topt} - 0.0005 \cdot \text{topt}^2$:
  1. $T_1 = \text{const} + \text{const} \cdot \text{topt} + \text{const} \cdot \text{topt}^2$
  2. $T_1 \rightarrow \text{const} | \text{const} + (T_1) \cdot \text{topt}$
CASA: a summary of the revision results

- relative accuracy improvement of about 9% (cross-validated)
- regarded non-trivial by scientists who developed the model
- taken into account in further model development
- reductions with single alternatives add up (orthogonal)
CAS4: the accuracy-similarity trade-off
CASA: the revised model

\[
NPP_c = \max(0, E \cdot IPAR)
\]

\[
E = 0.402 \cdot T1^{0.624} \cdot T2^{0.215} \cdot W^0
\]

\[
T1 = 0.680 + 0.270 \cdot topt - 0 \cdot topt^2
\]

\[
T2 = 1.1814/((1 + \exp(0.2 \cdot (TDIFF - 10))) \cdot (1 + \exp(0.3 \cdot (-TDIFF - 10))))
\]

\[
TDIFF = topt - tempc
\]

\[
TDIFF = topt - tempc
\]

\[
W = 0.5 + 0.5 \cdot eet/PET
\]

\[
PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet_tw_m
\]

\[
A = 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239
\]

\[
IPAR = FPAR_FAS \cdot \text{monthly_solar} \cdot SOL_CONV \cdot 0.5
\]

\[
FPAR_FAS = \min((SR_FAS - 1.08)/srdiff, 0.95)
\]

\[
SR_FAS = (1 + fas_ndvi/750)/(1 - fas_ndvi/750)
\]

\[
SOL_CONV = 0.0864 \cdot \text{days_per_month}
\]

\* note that the model is simpler than the initial one
inducing explanatory models from process-based knowledge: hipm
processes and models: representing explanations

- reconsider Lotka-Volterra model (ordinary differential equations)

\[
\frac{d}{dt} \text{hare} = 2.5 \cdot \text{hare} - 0.3 \cdot \text{hare} \cdot \text{fox}
\]

\[
\frac{d}{dt} \text{fox} = 0.1 \cdot 0.3 \cdot \text{hare} \cdot \text{fox} - 1.2 \cdot \text{fox}
\]

- three processes

  process \text{exponential\_growth} \quad \frac{d}{dt} \text{hare} = 2.5 \cdot \text{hare}

  process \text{exponential\_loss} \quad \frac{d}{dt} \text{fox} = -1.2 \cdot \text{fox}

  process \text{predator\_prey\_interaction} \quad \frac{d}{dt} \text{hare} = -0.3 \cdot \text{hare} \cdot \text{fox}

\quad \frac{d}{dt} \text{fox} = 0.1 \cdot 0.3 \cdot \text{hare} \cdot \text{fox}
generic vs. specific processes: representing knowledge

- generic process of predator prey interaction:
  
  process predator_prey_interaction
  
  variables\( Prey\{\text{species}\}, Pred\{\text{species}\} \)

  parameters\( r[0, \text{inf}], e[0, \text{inf}] \)

  equations
  \[
  \frac{d}{dt} Prey = -1 \cdot r \cdot Prey \cdot Pred \\
  \frac{d}{dt} Pred = e \cdot r \cdot Prey \cdot Pred
  \]

- instantiation \( Prey: \text{hare}, Pred: \text{fox}, r: 0.1, e: 0.3 \)

- specific (instantiated) process:
  
  process predator_prey_interaction

  
  \[
  \frac{d}{dt} hare = -0.3 \cdot hare \cdot fox \\
  \frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox
  \]
inducing process-based models: task definition

• GIVEN:
  – a set of generic processes
  – a set of system variables $V = \{v_1, v_2, \ldots v_n\}$
  – a table of measurements (observations) of the system variables

<table>
<thead>
<tr>
<th></th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$\ldots$</th>
<th>$v_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{1,1}$</td>
<td>$v_{2,1}$</td>
<td>$v_{n,1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$v_{1,2}$</td>
<td>$v_{2,2}$</td>
<td>$v_{n,2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ldots$</td>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$v_{1,m}$</td>
<td>$v_{2,m}$</td>
<td>$v_{n,m}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• FIND model
  – represented as a set of instantiated generic processes and
  – minimizes the discrepancy between measurements and simulation
problems of flat model representation: implausible models

- missing required/mandatory processes:
  - each species has a single growth process
  - each limiting factor has a growth limitation process

- mutually exclusive processes:
  - can not have two growth limitation processes
  - can not have two interaction processes between a pair of species

- solution:
  - hierarchical representation of generic processes
  - tells how to properly combine them into models
hierarchy of generic processes: population dynamics

- node legend: process, process type, optional (ellipse shaped)
- edge legend: AND/process (solid), OR/type (dashed)
hipm: inducing hierarchical process-based models

- search space
  - hierarchical process-based models
  - each represent a single equation structure

- beam search strategy
  1. start with structures that do not contain optionals
  2. evaluate structures (parameter estimation)
  3. place evaluated structures in the beam
  4. if the beam has not changed, stop and report the best model(s)
  5. expand/refine each structure on the beam
  6. goto second step
experiments with synthetic data: the methodology

- reconstruction of known models:
  - five simulation traces generated from random initial states
  - Gaussian noise added at 5% relative level

- three criteria for evaluating the performance:
  1. number of candidate models considered
  2. simulation error ($r^2$) of the induced models (cross-validated)
experiments with synthetic data: a summary of results

<table>
<thead>
<tr>
<th>bw</th>
<th>lib</th>
<th>#cms</th>
<th>CV-test $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>H</td>
<td>57.0 ± 1.6</td>
<td>0.99 ± 0.01</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>416.8 ± 18.4</td>
<td>0.93 ± 0.16</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>99.0 ± 2.3</td>
<td>0.99 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>797.8 ± 27.5</td>
<td>0.91 ± 0.16</td>
</tr>
</tbody>
</table>

- number of candidate models drastically reduced
  - reduction factor 7-8

- avoid inappropriate structure or parameters
  - inappropriate structure can lead to more accurate
  - but UNSTABLE models
Ross sea: task description

- environment and measurement data
  - concentration of three variables observed through time
  - 188 data points
  - what are the limiting factors for phytoplankton growth?

- observed variables:
  - phyto (phytoplankton)
  - nitrate (inorganic nutrient)
  - light intensity (estimated from ice concentration)
Ross sea: simulation of the induced model

- limiting factors: light and nitrate
- phosphate unimportant for phytoplankton growth
chemical kinetics of glycosis reaction network

- inducing (partial) chemical network of glycosis
  - determining causal connectivities of reacting species (compounds)
  - from temporal responses of species to pulse changes (14 time points)
  - Torralba et al. (2003) PNAS 100(4): 1494–1498

- responses of six chemical compounds measured over time:
  - G6P (glucose 6-phosphate)
  - F6P (fructose 6-phosphate)
  - F1,6BP (fructose 1,6-bisphosphate)
  - G3P (glycerol 3-phosphate)
  - 3PG (3-phosphoglycerate)
  - DHAP (dihydroxyacetone 3-phosphate)
glycosis: induced reaction network (sse = 452, r2 = 0.7940)

- similar to the one obtained by experts
  - "fuzzy" reasoning about dependencies
  - based on analysis of measured data
glycisis: simulation of the induced model (1)

![Graph showing observed and predicted values with RMSE and $r^2$ values for 3PG and F1,6BP.](image)

- **3PG**
  - Observed vs. Predicted Graph
  - RMSE = 2.61851, $r^2 = 0.9920$

- **F1,6BP**
  - Observed vs. Predicted Graph
  - RMSE = 2.48498, $r^2 = 0.9885$
glycosis: simulation of the induced model (2)

RMSE = 3.33982, $r^2 = 0.9942$

RMSE = 1.75741, $r^2 = 0.6923$
glycosis: simulation of the induced model (3)

![Graph showing the observed and predicted values for F6P and G6P with RMSE and $r^2$ values.]

RMSE = 1.73239, $r^2 = -0.0000$

RMSE = 5.85775, $r^2 = 0.0000$
summary, ongoing/further research, and open questions
summary

- CENTRAL ISSUE: transforming knowledge to inductive constraints
  - knowledge/model representation

- constrained space of candidate models beneficial
  - search space complexity reduction
  - error reduction
  - models can explain the observed phenomena
recent/ongoing research: scaletrack

- [Washio et al., 2005]: presented at this conference

- brief summary
  - hidden/unobserved variables: correlation dimensional analysis
  - improved parameter estimation: Bayesian framework
  - situations when no domain-specific knowledge is available
  - deals with chaotic dynamics (extremely sensitive models)
recent/ongoing research: bagging process-based models

- [Bridewell et al., 2005]: ICML
- brief summary
  - fusion of multiple models in a single one
  - to retain understandability
  - multiple models obtained on bootstrap data samples
  - improves convergence rate towards correct models
  - reduces overfitting
open issues

- alternative model quality measures
  - sum of squared errors sub-optimal in some domains
  - qualitative matching of time series
  - alternative parameter estimation techniques

- stricter constraints
  - only qualitative constraints for process combination
  - also quantitative cross-process parameter constraints

integration in interactive tools
SECOND PART
other computational discovery tasks
inference of regulatory chemical/genetic networks
**genepath: inference of genetic networks**

- [Zupan et al., 2003]: Bioinformatics

**GIVEN**
- experimental data from mutation (knock-out) experiments
gen-mutation-to-outcome relations
- knowledge in terms of partial network specification
gene-to-gene and gene-to-outcome relations

**FIND** one or more networks that are consistent with provided data and knowledge
genePath: two-step discovery

- first step — relating the genes with expert-defined patterns
  - influence: gene-to-outcome relationship (inhibits/activities)
  - parallel: both genes influence outcome, but not each other
  - epistatic: gene X precede gene Y in a path
  - no-influence: gene does not influence the outcome
output: set of gene-to-gene and gene-to-outcome relations

- expert can
  - review discovered relations
  - get the evidence for each relation

- second step — synthesis of the genetic network from
  - relations/constraints discovered in the first phase and
  - background knowledge (partial network specification)
genepath: qualitative simulation

- genepath capable of further network evaluation
  - qualitative simulation of the network
  - check simulation against observed data

- on-line demo: genepath.org
aggregation of dictystelium: task description

• experimental data
  – observed outcome: cell aggregation
  – influence of six genes on the outcome
  – 15 mutation experiments
  – each experiment reports the influence of gene mutation (knock out or overexpression) on outcome (no/reduced/normal/excessive aggregation)

• background knowledge — partial network specification
  – known influences of three genes on the cell aggregation
aggregation of dictyostelium: derived genetic network
inducing networks of chemical reactions with genetic programming

- [Koza et al., 2001-2003]
- brief summary
  - chemical network = set of chemical reactions
  - each reaction = function that relates substrates with products
  - functions re-presented as trees
  - appropriate cross-over operator defined
  - fitness measure: fit of simulation against data
  - only synthetic data experiments

computational scientific discovery (75)
inducing metabolic pathway networks

- [Garrett et al., 2001]

- brief summary
  - qualitative models of pathway networks
  - search through the space of qualitative models
  - qualitative simulation used to check model validity
  - background knowledge in terms of building blocks
  - example: glycolisis pathway
  - only synthetic data experiments reported
literature-based discovery
literature-based discovery: Swanson’s model

- GIVEN that scientific community published results on
  - A (substance) affects B (symptom)
  - B (symptom) is one of the characteristics of C (disease)

- ESTABLISH the hidden connection between A and C
  - e.g., substance A cures disease C
Swanson’s model: examples

- following the presented model Swanson’s established two previously unknown connections

- fish oil therapy of Raynaud’s disease by studies of:
  - effects of fish oil on (reduced) blood viscosity
  - Raynaud’s disease patients have high blood viscosity

- relation between magnesium insufficiency and migraine
  - based on 11 indirect connections published in literature

- two discoveries made manually
  - latter, arrowsmith method implemented
improved model for drug discovery

• by linking:
  – drug profile (A affects number of symptoms B) with
  – disease profile (C has number of symptoms B)

establish unknown relation between A and C
literature-based discovery: open vs. closed approach

• arrowsmith support closed discovery:
  – concepts A and C are both provided by the user
  – find hidden relations by searching for B(s)

• more general open approach:
  – only A is given
  – support tools for searching Bs and C

• in both approaches relations established
  – text analysis of bibliographic databases
  – by using frequent terms co-occurrence patterns
  – and semantic (markup) information
  – Medline, UMLS (MESH keywords)
CONCLUSION
what is computational scientific discovery
computational scientific discovery: task definition

- computational methods for automating/supporting scientific activities

- historical perspective:
  - early research focused on reconstructing important episodes from history of science
  - current research focus on solving current scientific problems and lead to novel scientific discoveries
computational scientific discovery: focus and emphasis

- focus on knowledge representation...
  - standard scientific formalisms
    equations, reaction pathways, structural models, etc.
  - invented and routinely used by scientists
- and not methods
  - variety of methods used with emphasis on heuristic search
  - other AI/ML/DM methods used too (bagging, gp, qr)
- obtained results should be
  - understandable/communicable with scientists in the field
  - publishable in relevant scientific literature
computational scientific discovery vs. data mining

- focus on methods
  - efficient/scalable methods for vast amounts of data
  - usual focus on business data or in data-rich scientific fields
    bioinformatics, remote sensing/imaging data
  - and not relevant knowledge representation formalisms
    - not routinely used by scientists, e.g. decision trees, rules
    - or frequent patterns (alternative data representation)
- results seldomly communicable to scientists in the field
- knowledge rich vs. data rich domains
computational scientific discovery: incremental activity

- build on existing knowledge or models:
  - methods for inducing new model/theories but also
  - methods for revision of existing ones

- integration of existing knowledge of primary importance

- interactive systems:
  - that support incremental aspects of establishing models/theories
BIBLIOGRAPHY
edited volumes


survey/overview articles


articles on equation discovery


articles on inducing models of dynamic systems

articles on genetic networks inference


articles on literature-based discovery


web sites

  Scientific Discovery AITopics @ aaai.org

- http://www-ai.ijs.si/ ljupco/ed/
  Equation Discovery

- http://www.isle.org/process.html
  Process-Based Modeling

- http://genepath.org/
  GenePath Demo
ACKNOWLEDGMENTS
thanks to/research teams

• lagrange and population dynamics applications:
  – Sašo Džeroski, Jožef Stefan Institute, Slovenia
  – Nataša Atanasova and Boris Kompare
    Faculty of civil engineering and geodesy, Slovenia

• hipm:
  – Pat Langley, Will Bridewell, and Oren Shiran
    Stanford University, USA

• casa:
  – Christopher Potter and Stephen Clooster
    NASA-Ames research center, USA

• some slides based on talks prepared by Will Bridewell, Pat Langley,
  and Sašo Džeroski