Need for High-Performance Computing in Analysis of Biomedical Data

Lee M. Hively and Vladimir Protopopescu
Oak Ridge National Laboratory
Oak Ridge, TN

Cray Biology Workshop
9 May 2003
High-Performance Computing (HPC)

- Increasing need for …
- Rapid optimization of medical device
- Real-time transmission, analysis, and assessment of clinical patient data
- Advanced visualization of patient data
Outline

• Detection of condition change
• Nonlinear measures
• Validation on model data
• Application to biomedical data
  - Epileptic seizures from scalp EEG
  - Cardiac events from surface ECG
  - Breathing difficulty from lung sounds
  - Sepsis onset from rat ECG
• Need for HPC in biomedical applications
• Conclusions
Detection of Condition Change

- Change of underlying parameter leads to ...
- Change in dynamics, which eventually ...
- Causes failure, disruption, or abnormal event

The challenge is to measure, detect, and assess change in nonlinear features timely, robustly, and accurately from time serial data

Assumptions: deterministic underlying dynamics
\[ \tau(\text{data}) \gg \tau(\text{dynamics}) \]
known origin of artifacts
Linear Time Series Analysis

1. Discretely sampled single- or multi-channel data:
   \[ x_i^k = x^k(t_i), \ i = 1, 2, \ldots, N, \text{ with } t_i = i \Delta t \]
   \[ k = 1, \ldots, K; \quad \text{for } K \text{ channels} \]

2. Remove noise and/or artifact via some filtering technique

3. Apply signal analysis of choice: Fourier, wavelets, SVD, etc.

4. Define various measures: power spectra, correlation functions, statistical moments, etc.
Nonlinear Time Series Analysis

1. Single-channel time serial data, as before

2. Time-delayed vector $y(i) = [x_i, x_{i+\lambda}, ..., x_{i+(d-1)\lambda}]$ lives in $d$-dimensional phase space (PS): $d$ and $\lambda$ not known \textit{a priori}, but can be found.

3. Non-transient $y(i)$ evolves on a fractal attractor with $d-1 \leq D \leq d$

4. Discretization determines an invariant distribution function (DF)
   
   $P_k =$ occurrence frequency in $k$-th discrete bin

   Unchanging DF $\iff$ unaltered dynamics
   Changing DF $\iff$ condition change (new $P_k$ or new PS location)

5. Compare baseline ($P_k$) and test ($Q_k$) DFs via nonlinear measures
Phase Space Dissimilarity Measures

- $\chi^2 = \sum_k (P_k - Q_k)^2/(P_k + Q_k)$
- $L = \sum_k |P_k - Q_k|$
- New measures are more discriminating than traditional ones since:
  - Dissimilarity measures: subtract, then integrate … more sensitive
  - Traditional measures: integrate, then subtract … less sensitive
Validation on the Lorenz Model

- Three (3) simultaneous ODEs:
  \[ \frac{dx}{dt} = a(y - x) \]
  \[ \frac{dy}{dt} = rx - y - xz \]
  \[ \frac{dz}{dt} = xy - bx \]
- Parameters: \( a=10, b=8/3, 25 \leq r \leq 90 \)
- Basecase for \( r=25 \)
- 200,000 points for each value of \( r \)
  four non-overlapping sets of 50,000 points
Phase Space (PS) for Lorenz Model
Results for Lorenz Model
Reliable Analysis Needs Long Cutsets

Long Cutsets Require HPC
Application to EEG Data

- Biomedical Monitoring Systems Inc.
- Sample rate = 250 Hz
- 19 channels of scalp data
- Band-pass filtered between 0.5 and 99 Hz
- Datasets span 5000 – 29500s
- 60 datasets: 40 event and 20 non-event
- Multiple datasets: 30 from 11 patients
- 36 females and 24 males: $4 \leq \text{age} \leq 57$
Condition Change in EEG
Application to ECG Data

- ORNL evaluated 5 EKG datasets
  - Digital data from Holter recordings
  - Analysis of one channel, sampled at 250 Hz
  - Datasets spanned < 1 hour
- PS-DF provides > 6 minutes of forewarning
  - No false positives or negatives in 5 datasets
  - Need more data for more reliable validation
Fibrillation Forewarning from ECG

(a) ECG (AU)

(b) U(t)

(c) U(K)

(d) U(t)^2

(e) FOREWARNING TIME (SEC) = 960
Breathing difficulty from lung sounds

- Experiment at Walter Reed Medical Ctr.
  - anesthetized pig
  - 0 – 1400 ml of air into pleural space
  - surface stethoscope
  - sampling rate = 10 kHz
- Basecase for normal breathing (0 ml)
- Testcases for 100 ml increments
Example of Results for Pig Lung Sounds
ECG Analysis for Sepsis Onset

- 23 anesthetized rats at UTKMC
  - 17 exposed to endotoxin via inhalation
  - 6 with no exposure (de-ionized water)
- 4 surface ECG electrodes
  - sample rate = 500 Hz
- Experimental protocol (1.5 - 3 hours total)
  - 30-60 minutes for acclimation (basecase)
  - 30 minutes of *Salmonella* endotoxin
  - 30-90 minutes of recovery
Rat Sepsis Results: Total True Rate = 1
Vision for Advanced Biomedical Analysis

- New sensors will provide ...
- Wealth of data
  - sounds, EEG, ECG, blood, images, etc. for ...
- Fusion of multi-channel data from ...
- Dynamics of metabolic/sensory networks with ...
- Near-real-time response, yielding ...
- New medical science and ...
- Lower cost, higher quality care of patients
Compelling Need for HPC

- HP network transmission of real-time data
- Many data channels (10s to 1000s to millions)
- HP data storage
- HP analysis (parallelize over channels)
- Advanced visualization
Acknowledgments

- Sponsors:
  - ORNL LDRD (Seed Money): EEG, sepsis
  - USDOE OEM, BES, OS/LTR
  - Nicolet Biomedical Inc. (now ViaSys Healthcare Inc.)
    plus EEG data and characterization
- PhysioControls Inc. – ECG data
- Ned Clapp (ORNL – retired)
- Glenn Allgood (ORNL) – lung sounds data
- Dennis Strickler – initial code parallelization
- Nancy Munro (ORNL) – human physiology
- Kathe Fischer (ORNL) – code improvements
- Michael Karlstad, Walter Holland, Matt Vassy (UTMCK) – ECG/sepsis data and characterization
Backups
Work to Date

- Desktop 2GHz P4 Wintel PC (1GB memory)
  - Executable size \( \leq 55\text{MB} \)
  - Amount of data \( \sim 30\text{GB} \)
  - \( \leq 12\text{h} \) run to analyze 261h of EEG
    - \((0.5 \text{ CPU-d})/\text{run} \times (\text{manual optimization})=\text{months}\)
- First attempt at parallelization: non-scalable
  - huge input overhead
Traditional Nonlinear Measures

Shannon entropy: $e_P = \sum_k P_k \log P_k$

Dissimilarity: $e_P - e_Q$

$e_P$ and $e_Q$ may be close, even if $P_k$ and $Q_k$ are not!

Same for other traditional nonlinear measures:

- Kolmogorov entropy
- Correlation dimension
- Mutual information function
- Lyapunov exponents
- Fractal dimensions
- Recurrence times
Bondarenko Model

• Time-delayed set of nonlinear ODEs:
  \[ \frac{du_i}{dt} = -u_i + \sum_{i \neq k} a_{ik} f(u_k(t-\tau)) \quad i = 1,\ldots,M \]

• Parameters:
  -2 \leq a_{ik} \leq 2, \quad M = 10, \quad \tau = 10, \quad 5 \leq c \leq 18

• Function: \( f(x) = c \tanh(x) \)

• Basecase for \( c = 5 \)

• 100,000 points for each value of \( c \)
  four non-overlapping sets of 25,000 points
Results for Bondarenko Model