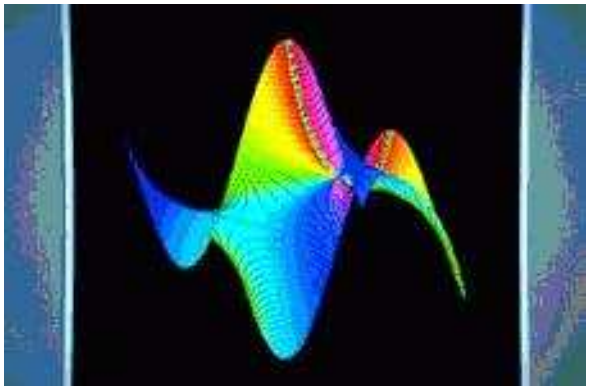
Information Transfer in Biological Systems*

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

July 1, 2007

AofA and **IT** logos



*Joint work with A. Grama, P. Jacquet, M. Koyoturk, M. Regnier, and G. Seroussi.

Outline

1. What is Information?
2. Beyond Shannon Information
3. Information Transfer: Darwin Channel
   - Model: Deletion and Constrained Channels
   - Capacity of the Noisy Constrained Channel
4. Information Discovery in Massive Data: Pattern Matching
   - Classification of Pattern Matching Problems
   - Statistical Significance
   - Finding Weak Signals in Biological Data
5. Information in Structures: Network Motifs
   - Biological Networks
   - Finding Biologically Significant Structures

What is Information?

C. F. Von Weizsäcker:

"Information is only that which produces information" (relativity).
"Information is only that which is understood" (rationality)
"Information has no absolute meaning.

C. Shannon:
"These semantic aspects of communication are irrelevant . . ."

F. Brooks, jr. (JACM, 50, 2003, "Three Great Challenges for . . . CS "):

"Shannon and Weaver performed an inestimable service
by giving us a definition of Information and a metric for
for Information as communicated from place to place.
We have **no theory** however that gives us a metric
for the Information embodied in **structure** . . .
this is the most fundamental gap in the theoretical underpinning of Information and computer science. . . . A young information theory scholar willing to spend years on a deeply fundamental problem need look no further."

Some Definitions

Information has flavor of:

relativity (depends on the activity undertaken),
rationality (depends on the recipient's knowledge),
timeliness (temporal structure),
space (spatial structure).

Informally Speaking:
A piece of data carries information if it can impact a recipient's ability to achieve the objective of some activity within a given context.

Definition 1. *The amount of information (in a faultless scenario)* $\mathrm{info}(E)$ *carried by the event* E *in the context* C *as measured for a system with the rules of conduct* R *is*

$$\mathrm{info}_{R,C}(E) = \mathbf{cost}[\mathrm{objective}_R(C(E)), \mathrm{objective}_R(C(E) + E)]$$

where the **cost** *(weight, distance) is taken according to the ordering of points in the space of objectives.*

Shannon Information Theory

In our setting, Shannon defined:

objective: statistical ignorance of the recipient;
statistical uncertainty of the recipient.

cost: # binary decisions to describe E;
$= -\log P(E)$; $P(E)$ being the probability of E.

Context: the semantics of data is irrelevant . . .

Self-information for E_i: $\text{info}(E_i) = -\log P(E_i)$.
Average information: $H(P) = -\sum_i P(E_i) \log P(E_i)$
Entropy of $X = \{E_1, \ldots\}$: $H(X) = -\sum_i P(E_i) \log P(E_i)$
Mutual Information: $I(X;Y) = H(Y) - H(Y|X)$, (faulty channel).

Theorem 2. (**Shannon 1948; Channel Coding**)
In Shannon's words:

It is possible to send information at the capacity through the channel with as small a frequency of errors as desired by proper (**long**) encoding. This statement is not true for any rate greater than the capacity.

(The maximum codebook size $N(n, \varepsilon)$ for codelength n and error probability ε is asymptotically equal to: $N(n, \varepsilon) \sim 2^{nC}$.)

Information in Biology

M. Eigen

"The differentiable characteristic of the living systems is Information. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability . . . . Information theory, pioneered by **Claude Shannon**, cannot answer this question . . . in principle, the answer was formulated 130 years ago by **Charles Darwin**.

Some Fundamental Questions:

- how information is generated and transferred through underlying mechanisms of variation and selection;
- how information in biomolecules (sequences and structures) relates to the organization of the cell;
- whether there are error correcting mechanisms (codes) in biomolecules;
- and how organisms survive and thrive in noisy environments.

Life is a delicate interplay of energy, entropy, and information; essential functions of living beings correspond to the generation, consumption, processing, preservation, and duplication of **information**.

Beyond Shannon

Participants of the **2005 Information Beyond Shannon** workshop realize:

Time: When information is transmitted over networks of gene regulation, protein interactions, the associated delay is an important factor.
(e.g., timely information exchange in cells may be responsible for bidirectional microtubule-based transport in cells).

Space: In molecular interaction networks, spatially distributed components raise fundamental issues of limitations in information exchange since the available resources must be shared, allocated, and re-used.

Information and Control: Again in networks involving regulation, signaling, and metabolism, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective.

Semantics: In many contexts, experimentalists are interested in signals, without precise knowledge of what they represent.

Dynamic information: In a complex network in a space-time-control environment (e.g., human brain information is not simply communicated but also processed) how can the consideration of such dynamical sources be incorporated into the Shannon-theoretic model?

Outline Update

1. What is Information?
2. Beyond Shannon Information
3. Information Transfer: Darwin Channel
   - Model: Deletion and Constrained Channels
   - Capacity of the Noisy Constrained Channel
4. Information Discovery in Massive Data: Pattern Matching
5. Information in Structures: Network Motifs

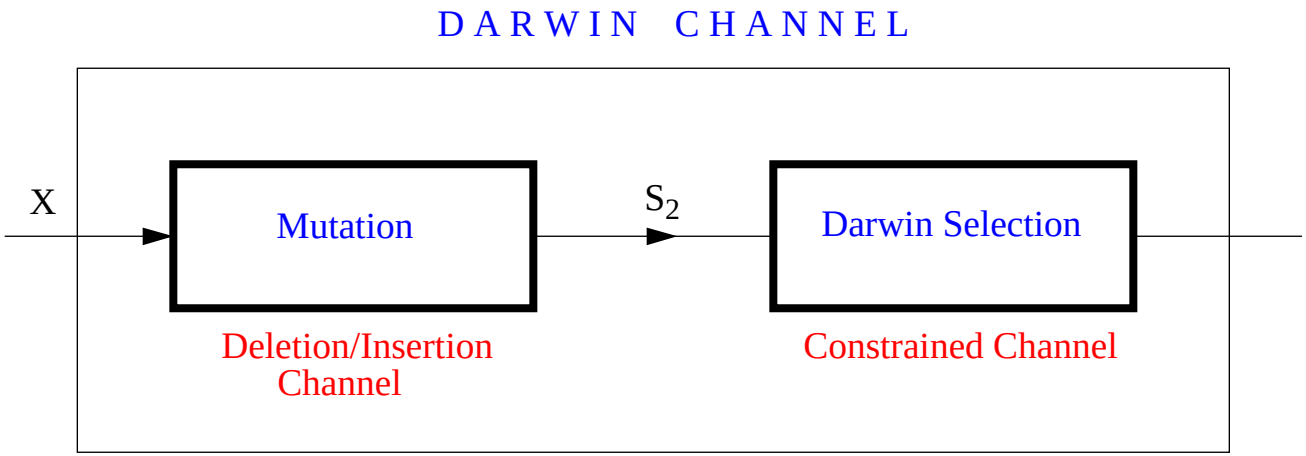
Darwin Channel

Biomolecular structures, species, and in general biodiversity, have gone through significant metamorphosis over eons through mutation and natural selection, which we model by constrained sequences/channels.

To capture mutation and natural selection we introduce

Darwin channel

which is a combination of a **noisy deletion/insertion channel** and a **noisy constrained channel**.



Noisy Constrained Channel

1. Binary Symmetric Channel (BSC):
(i) crossover probability ε,
(ii) **constrained set of inputs** (Darwin preselected) that can be modeled by a **Markov Process**,
(ii) $\mathcal{S}_n$ denotes the set of binary constrained sequences of length n.

2. Channel Input and Output:
Input: Stationary process $X = \{X_k\}_{k\geq 1}$ supported on $\mathcal{S} = \bigcup_{n>0} \mathcal{S}_n$.
Channel Output: **Hidden Markov Process** (HMP)

$$Z_i = X_i \oplus E_i$$

where $\oplus$ denotes addition modulo 2, and $E = \{E_k\}_{k\geq 1}$, independent of X, with $P(E_i = 1) = \varepsilon$ is a Bernoulli process (noise).

Note: To focus, we illustrate our results on

$$\mathcal{S}_n = \{\text{(d,k) sequences}\}$$

i.e., no sequence in $\mathcal{S}_n$ contains a run of zeros of length ishorter than d or longer than k. Such sequences can model neural spike trains (no two spikes in a short time).

Noisy Constrained Capacity

$C(\varepsilon)$ – conventional BSC channel capacity $C(\varepsilon) = 1 - H(\varepsilon)$, where $H(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$.

$C(\mathcal{S}, \varepsilon)$ – noisy constrained capacity defined as

$$C(\mathcal{S}, \varepsilon) = \sup_{X \in \mathcal{S}} I(X; Z) = \lim_{n \to \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Z_1^n),$$

where the suprema are over all stationary processes supported on $\mathcal{S}$ and $\mathcal{S}_n$, respectively. **This is an open problem since Shannon.**

Mutual information

$$I(X; Z) = H(Z) - H(Z|X)$$

where $H(Z|X) = H(\varepsilon)$.

Thus, we must find the entropy $H(Z)$ of a **hidden Markov process**! (e.g., (d, k) sequence can be generated as an output of a kth order Markov process).

Hidden Markov Process

1. Let $X = \{X_k\}_{k\geq 1}$ be a rth order stationary Markov process over a binary alphabet $\mathcal{A}$ with transition probabilities $P(X_t = a|X_{t-r}^{t-1} = a_1^r)$, where $a_1^r \in \mathcal{A}^r$. For $r = 1$, with transition matrix $\mathbf{P} = \{p_{ab}\}_{a,b\in\{01,\}}$.

2. Let $\bar{X} = 1 \oplus X$. In particular, $Z_i = X_i$ if $E_i = 0$ and $Z_i = \bar{X}_i$ if $E_i = 1$:

$$\begin{aligned} P(Z_1^n, E_n) &= P(Z_1^n, E_{n-1} = 0, E_n) + P(Z_1^n, E_{n-1} = 1, E_n) \\ &= P(Z_1^{n-1}, Z_n, E_{n-1} = 0, E_n) + P(Z_1^{n-1}, Z_n, E_{n-1} = 1, E_n) \\ &= P(E_n)P_X(Z_n \oplus E_n|Z_{n-1})P(Z_1^{n-1}, E_{n-1} = 0) \\ &\quad + P(E_n)P_X(Z_n \oplus E_n|\bar{Z}_{n-1})P(Z_1^{n-1}, E_{n-1} = 1) \end{aligned}$$



Entropy as a Product of Random Matrices

Let

$$\mathbf{p}_n = [P(Z_1^n, E_n = 0), P(Z_1^n, E_n = 1)]$$

and

$$\mathbf{M}(Z_{n-1}, Z_n) = \begin{bmatrix} (1-\varepsilon)P_X(Z_n|Z_{n-1}) & \varepsilon P_X(\bar{Z}_n|Z_{n-1}) \\ (1-\varepsilon)P_X(Z_n|\bar{Z}_{n-1}) & \varepsilon P_X(\bar{Z}_n|\bar{Z}_{n-1}) \end{bmatrix}.$$

Then

$$\mathbf{p}_n = \mathbf{p}_{n-1}\mathbf{M}(Z_{n-1}, Z_n).$$

and

$$P(Z_1^n) = \mathbf{p}_1\mathbf{M}(Z_1, Z_2)\cdots\mathbf{M}(Z_{n-1}, Z_n)\mathbf{1}^t,$$

where $\mathbf{1}^t = (1, \ldots, 1)$.

Thus, $P(Z_1^n)$ is a product of random matrices since $P_X(Z_i|Z_{i-1})$ are random variables.

Entropy Rate as a Lyapunov Exponent

Theorem 1 (Furstenberg and Kesten, 1960). *Let* $\mathbf{M}_1, \ldots, \mathbf{M}_n$ *form a stationary ergodic sequence and* $\mathbf{E}[\log^+ ||\mathbf{M}_1||] < \infty$ *Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}[\log ||\mathbf{M}_1 \cdots \mathbf{M}_n||] = \lim_{n\to\infty} \frac{1}{n} \log ||\mathbf{M}_1 \cdots \mathbf{M}_n|| = \mu \quad \text{a.s.}$$

where μ *is called top Lyapunov exponent.*

Corollary 1. *Consider the HMP* Z *as defined above. The entropy rate*

$$\begin{aligned} h(Z) &= \lim_{n\to\infty} \mathbf{E}[\, -\frac{1}{n} \log P(Z_1^n)\,] \\ &= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}[-\log\left(\mathbf{p}_1 \mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n)\mathbf{1}^t\right)] \end{aligned}$$

is a top Lyapunov exponent of $\mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n)$.

Unfortunately, it is notoriously difficult to compute top Lyapunov exponents as proved in Tsitsiklis and Blondel. Therefore, in next we derive an explicit asymptotic expansion of the entropy rate $h(Z)$.

Asymptotic Expansion

We now assume that $P(E_i = 1) = \varepsilon \to 0$ is small (e.g., $\varepsilon = 10^{-12}$ for mutation).

Theorem 2 (Seroussi, Jacquet and W.S., 2004). *Assume rth order Markov. If the conditional probabilities in the Markov process X satisfy*

$$P(a_{r+1}|a_1^r) > 0 \quad \textbf{IMPORTANT!}$$

for all $a_1^{r+1} \in \mathcal{A}^{r+1}$, then the entropy rate of Z for small ε is

$$h(Z) = \lim_{n\to\infty} \frac{1}{n} H_n(Z^n) = h(X) + f_1(P)\varepsilon + O(\varepsilon^2),$$

where

$$f_1(P) = \sum_{z_1^{2r+1}} P_X(z_1^{2r+1}) \log \frac{P_X(z_1^{2r+1})}{P_X(\bar{z}_1^{2r+1})} = \mathbb{D}\left(P_X(z_1^{2r+1}) || P_X(\bar{z}_1^{2r+1})\right),$$

where $\bar{z}^{2r+1} = z_1 \ldots z_r \bar{z}_{r+1} z_{r+2} \ldots z_{2r+1}$. In the above, $h(X)$ is the entropy rate of the Markov process X, $\mathbb{D}$ denotes the Kullback-Liebler divergence.

Examples

Example 1. Consider a Markov process with symmetric transition probabilities $p_{01} = p_{10} = p$, $p_{00} = p_{11} = 1-p$. This process has stationary probabilities $P_X(0) = P_X(1) = \frac{1}{2}$. Then

$$h(Z) = h(X) + f_1(p)\varepsilon + f_2(p)\varepsilon^2 + O(\varepsilon^3)$$

where

$$f_1(p) = 2(1-2p)\log\frac{1-p}{p}, \qquad f_2(p) = -f_1(p) - \frac{1}{2}\left(\frac{2p-1}{p(1-p)}\right)^2 .$$

Example 2. (Degenerate Case.) Consider the following Markov process

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}$$

where $0 \le p \le 1$.
Ordentlich and Weissman (2004) proved for this case

$$H(Z) = H(P) - \frac{p(2-p)}{1+p}\varepsilon\log\varepsilon + O(\varepsilon)$$

(e.g., $(11\ldots)$ will not be generated by MC, but can be outputed by HMM with probability $O(\varepsilon^\kappa)$).

Exact Noisy Constrained Capacity

Recall $I(X;Z) = H(Z) - H(\varepsilon)$. Then by above theorem

$$H(Z) = \mu(P)$$

where $\mu(P)$ is the top Lyapunov exponent of $\{\mathbf{M}(\widetilde{Z}_i|\widetilde{Z}_{i-1})\}_{i>0}$.

It is known (cf. Chen and Siegel, 2004) that he process optimizing the mutual information can be approached by a sequence of Markov probabilities $P^{(r)}$ of the constraint of increasing order.

Theorem 3. *The noisy constrained capacity $C(\mathcal{S},\varepsilon)$ for a (d,k) constraint through a BSC channel of parameter ε is given by*

$$C(\mathcal{S},\varepsilon) = \lim_{r\to\infty} \sup_{P^{(r)}} \mu(P^{(r)}) - H(\varepsilon)$$

where $P^{(r)}$ denotes the probability law of an rth-order Markov process generating the (d,k) constraint $\mathcal{S}$.

Main Asymptotic Results

We observe (cf. Han and Marcus (2007))

$$H(Z) = H(P) - f_0(P)\varepsilon \log \varepsilon + f_1(P)\varepsilon + o(\varepsilon)$$

for explicitly computable $f_0(P)$ and $f_1(P)$.

Let $P^{\max}$ be the maxentropic maximizing $H(P)$. Then

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) - (1 - f_0(P^{\max}))\varepsilon \log \varepsilon + (f_1(P^{\max}) - 1)\varepsilon + o(\varepsilon)$$

where $C(\mathcal{S})$ is known capacity of a noiseless channel.

Example: For (d, k) sequences, we can prove:

(i) for $k \leq 2d$

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) + A \cdot \varepsilon + O(\varepsilon^2 \log \varepsilon)$$

(ii) For $k > 2d$

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) + B \cdot \varepsilon \log \varepsilon + O(\varepsilon),$$

where $A \& B$ are computable constants (cf. also Han and Marcus (2007)).

Outline Update

1. What is Information?
2. Information Transfer: Darwin Channel
3. Information Discovery in Massive Data: Pattern Matching
   - Classification of Pattern Matching Problems
   - Statistical Significance
   - Finding Weak Signals in Biological Data
4. Information in Structures: Network Motifs

Pattern Matching

Let $\mathcal{W} = w_1 \ldots w_m$ and T be strings over a finite alphabet $\mathcal{A}$.

Basic question: **how many times $\mathcal{W}$ occurs in T**.

Define $O_n(\mathcal{W})$ — the number of times $\mathcal{W}$ occurs in T, that is,

$$O_n(\mathcal{W}) = \#\{i : \ T^i_{i-m+1} = \mathcal{W}, \ m \leq i \leq n\}.$$

Generalized String Matching: A set of patterns is given, that is,

$$\mathcal{W} = (\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_d), \quad \mathcal{W}_i \in \mathcal{A}^{m_i}$$

where $\mathcal{W}_i$ itself for $i \geq 1$ is a subset of $\mathcal{A}^{m_i}$ (i.e., a set of words of a given length m_i). The set $\mathcal{W}_0$ is called the forbidden set.

Three cases to be considered:

$\mathcal{W}_0 = \emptyset$ — the number of patterns from $\mathcal{W}$ occurring in the text.

$\mathcal{W}_0 \neq \emptyset$ — the number of $\mathcal{W}_i$, $i \geq 1$ pattern occurrences under the condition that no pattern from $\mathcal{W}_0$ occurs in the text.

$\mathcal{W}_i = \emptyset$, $\imath \geq 1$, $\mathcal{W}_0 \neq \emptyset$ — e.g., (d, k) sequences.

Probabilistic Sources

Memoryless Source

The text is a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol $s \in \mathcal{A}$ occurs with probability $P(s)$.

Markovian Source

The text is a realization of a stationary Markov sequence of order r, that is, probability of the next symbol occurrence depends on r previous symbols.

Basic Thrust of our Approach

When searching for over-represented or under-represented patterns we must assure that such a pattern is not generated by randomness itself (to avoid too many false positives).

Z Score vs p-values

Some statistical tools are used to characterize underrepresented and overrepresented patterns. We illustrate it on $O_n(\mathcal{W})$.

Z-scores

$$Z(\mathcal{W}) = \frac{\mathbf{E}[O_n] - O_n(\mathcal{W})}{\sqrt{\mathbf{Var}[O_n(\mathcal{W})]}}$$

i.e., how many standard deviations the observed value $O_n(\mathcal{W})$ is away from the mean.

This score makes sense only if one can prove that Z satisfies (at least asymptotically) the Central Limit Theorem (CLT), that is, Z is normally distributed.

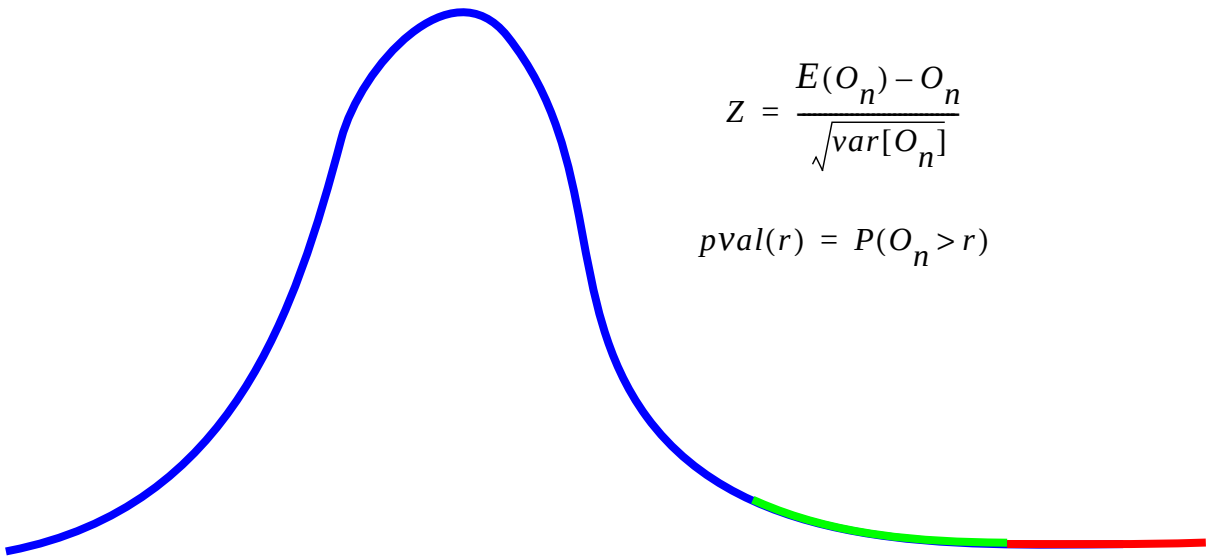
p-values

$$pval(r) = P(O_n(\mathcal{W}) > \underbrace{\mathbf{E}[O_n] + x\sqrt{\mathbf{Var}[O_n]}}_{r});$$

p values are used for very rare occurrences, far away from the mean.

In order to compute p values one must apply either Moderate Large deviation (MLD) or Large Deviations (LD) results.

CLT vs LD



Let

$$P(O_n \geq n\alpha + x\sigma\sqrt{n})$$

Central Limit Theorem (CLT) – valid only for $x = O(1)$.

Moderate Large Deviations (MLD) – valid for $x \to \infty$ but $x = o(\sqrt{n})$.

Large Deviations (MLD) – valid for $x = O(\sqrt{n})$.

Z-scores and p values for A.thaliana

Table 1: Z score vs p-value of tandem repeats in *A.thaliana*.

Oligomer	Obs.	p-val (large dev.)	Z-sc.
AATTGGCGG	2	8.059×10^{-4}	48.71
TTTGTACCA	3	4.350×10^{-5}	22.96
ACGGTTCAC	3	2.265×10^{-6}	55.49
AAGACGGTT	3	2.186×10^{-6}	48.95
ACGACGCTT	4	1.604×10^{-9}	74.01
ACGCTTGG	4	5.374×10^{-10}	84.93
GAGAAGACG	5	0.687×10^{-14}	151.10

Remark: p values were computed using large deviations results of Regnier and W.S. (1998), and Denise and Regnier (2001) as we discuss below.

Some Theory: Language Based Approach

Here is an incomplete list of results on string pattern matching:
A. Apostolico, Feller (1968), Prum, Rodolphe, and Turckheim, (1995), Regnier & W.S. (1997,1998), P. Nicodéme, Salvy, & P. Flajolet (1999).

Language Approach: Language $\mathcal{L}$ a collection of words, and its generating function is

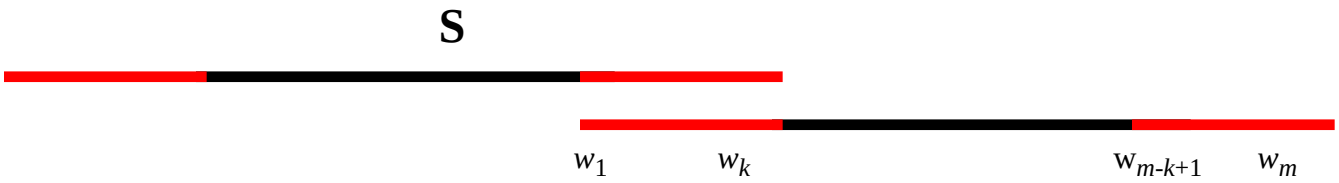
$$L(z) = \sum_{u \in \mathcal{L}} P(u) z^{|u|}$$

where $P(u)$ is the probability u occurrence, $|u|$ is the length of u.

Autocorrelation Set and Polynomial:
Given a pattern $\mathcal{W}$, we define the autocorrelation set $\mathcal{S}$ as:

$$\mathcal{S} = \{w_{k+1}^m : \ w_1^k = w_{m-k+1}^m\}, \quad w_1^k = w_{m-k+1}^m$$

and $\mathcal{WW}$ is the set of positions k satisfying $w_1^k = w_{m-k+1}^m$.



The generating function of $\mathcal{S}$ is denoted as $S(z)$ and we call it the autocorrelation polynomial.

$$S(z) = \sum_{k \in \mathcal{WW}} P(w_{k+1}^m) z^{m-k}.$$

Language $\mathcal{T}_r$

$\mathcal{T}_r$ – set of words that contains exactly $r \geq 1$ occurrences of $\mathcal{W}$. Define

$$T_r(z) = \sum_{n \geq 0} \Pr\{O_n(\mathcal{W}) = r\} z^n, \quad T(z,u) = \sum_{r=1}^{\infty} T_r(z) u^r.$$

(i) We define $\mathcal{R}$ as the set of words containing only one occurrence of $\mathcal{W}$, located at the right end; e.g., , for $\mathcal{W} = aba$, $ccaba \in \mathcal{R}$.

(ii) We also define $\mathcal{U}$ as

$$\mathcal{U} = \{u : \ \mathcal{W} \cdot u \cdot \in \mathcal{T}_1\}$$

that is, a word $u \in \mathcal{U}$ if $\mathcal{W} \cdot u$ has exactly one occurrence of $\mathcal{W}$ at the left end of $\mathcal{W} \cdot u$; e.g., $bba \in \mathcal{U}$ but $ba \notin \mathcal{U}$.

(iii) Let $\mathcal{M}$ be the language:

$$\mathcal{M} = \{u : \ \mathcal{W} \cdot u \in \mathcal{T}_2 \text{ and } \mathcal{W} \text{ occurs at the right of } \mathcal{W} \cdot u\},$$

that is, $\mathcal{M}$ is a language such that $\mathcal{W}\mathcal{M}$ has exactly two occurrences of $\mathcal{W}$ at the left and right end of a word from $\mathcal{M}$; e.g., $ba \in \mathcal{M}$ since $ababa \in \mathcal{T}_2$.

Basic Lemma

Lemma 1. *The language $\mathcal{T}$ satisfies the fundamental equation:*

$$\mathcal{T} = \mathcal{R} \cdot \mathcal{M}^* \cdot \mathcal{U} \ .$$

Notably, the language $\mathcal{T}_r$ can be represented for any $r \geq 1$ as follows:

$$\mathcal{T}_r = \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U},$$

and

$$\mathcal{T}_0 \cdot \mathcal{W} = \mathcal{R} \cdot \mathcal{S} \ .$$

Here, by definition $\mathcal{M}^0 := \{\epsilon\}$ and $\mathcal{M}^* := \bigcup_{r=0}^{\infty} M^r$.

Example: Let $\mathcal{W} = TAT$. The following string belongs $\mathcal{T}_3$:

$$\overbrace{CCTAT}^{\mathcal{R}} \underbrace{AT}_{\mathcal{M}} \underbrace{GATAT}_{\mathcal{M}} \overbrace{GGA}^{\mathcal{U}} .$$

Main Results – Exact

Theorem 4. (i) *The languages $\mathcal{M}, \mathcal{U}$ and $\mathcal{R}$ satisfy:*

$$\begin{aligned} \bigcup_{k\geq 1} \mathcal{M}^k &= \mathcal{A}^* \cdot \mathcal{W} + \mathcal{S} - \{\epsilon\}\ , \\ \mathcal{U} \cdot \mathcal{A} &= \mathcal{M} + \mathcal{U} - \{\epsilon\}, \\ \mathcal{W} \cdot \mathcal{M} &= \mathcal{A} \cdot \mathcal{R} - (\mathcal{R} - \mathcal{W})\ , \end{aligned}$$

where $\mathcal{A}^*$ is the set of all words, $+$ and $-$ are disjoint union and subtraction of languages.

(ii) *The generating functions $T_r(z)$ and $T(z,u)$ are*

$$\begin{aligned} T_r(z) &= R(z)M^{r-1}(z)U(z), \quad r \geq 1 \\ T(z,u) &= R(z)\frac{u}{1-uM(z)}U(z) \\ T_0(z)P(\mathcal{W}) &= R(z)S(z) \end{aligned}$$

where

$$M(z) = 1 + \frac{z-1}{D(z)}, \quad U(z) = \frac{1}{D(z)}, \quad R(z) = z^m P(\mathcal{W})\frac{1}{D(z)}.$$

with $D(z) = (1-z)S(z) + z^m P(\mathcal{W})$.

Main Results: Asymptotics

Theorem 5. (i) *Moments. The expectation satisfies, for* $n \geq m$*:*

$$\mathbf{E}[O_n(\mathcal{W})] = P(\mathcal{W})(n - m + 1) \ ,$$

while the variance is

$$\mathbf{Var}[O_n(\mathcal{W})] = nc_1 + c_2.$$

where c_1, c_2 *are explicit constants.*

(ii) *CLT: Case* $r = EO_n + x\sqrt{\mathbf{Var}O_n}$ *for* $x = O(1)$*. Then:*

$$\Pr\{O_n(\mathcal{W}) = r\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \ .$$

(iii) *Large Deviations: Case* $r = (1 + \delta)EO_n$*. Let* $a = (1 + \delta)P(\mathcal{W})$*, then*

$$\Pr\{O_n(\mathcal{W}) \quad \sim \quad (1 + \delta)EO_n\} = \frac{e^{-(n-m+1)I(a)+\delta_a}}{\sigma_a\sqrt{2\pi(n - m + 1)}}$$

where $I(a) = a\omega_a + \rho(\omega_a)$ *and* $\delta_a, \omega_a, \rho(\omega_a)$ *are constants, and* ρ *is the smallest root of* $D(z) = 0$*.*

Biology – Weak Signals and Artifacts

Denise and Regnier (2002) observed that in biological sequence whenever a word is overrepresented, then its subwords and proximate words are also likely to be overrepresented (the so called artifacts).
Example: if $\mathcal{W}_1 = AATAAA$, then $\mathcal{W}_2 = ATAAAN$ is also overrepresented.

New Approach:

> Once a dominating signal has been detected, we look for a weaker signal by comparing the number of observed occurrences of patterns to the conditional expectations **not** the regular expectations.

In particular, using the methodology presented above Denise and Regnier (2002) were able to prove that

$$\mathbf{E}[O_n(\mathcal{W}_2)|O_n(\mathcal{W}_1) = k] \sim \alpha n$$

provided $\mathcal{W}_1$ is overrepresented, where α can be explicitly computed (often $\alpha = P(\mathcal{W}_2)$ is $\mathcal{W}_1$ and $\mathcal{W}_2$ do not overlap).

Polyadenylation Signals in Human Genes

Beaudoing et al. (2000) studied several variants of the well known AAUAAA polyadenylation signal in mRNA of humans genes. To avoid artifacts Beaudoing et al cancelled all sequences where the overrepresented hexamer was found.

Using our approach Denise and Regnier (2002) discovered/eliminated all artifacts and found new signals in a much simpler and reliable way.

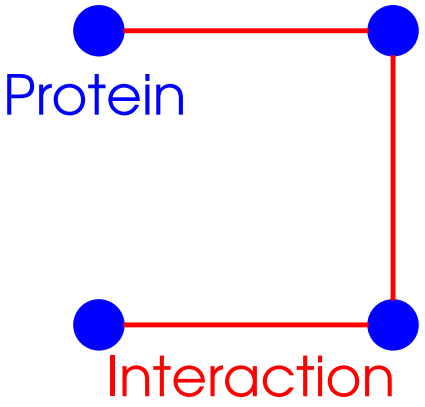
Hexamer	Obs.	Rk	Exp.	Z-sc.	Rk	Cd.Exp.	Cd.Z-sc.	Rk
AAUAAA	3456	1	363.16	167.03	1			1
AAAUAA	1721	2	363.16	71.25	2	1678.53	1.04	1300
AUAAAA	1530	3	363.16	61.23	3	1311.03	6.05	404
UUUUUU	1105	4	416.36	33.75	8	373 .30	37.87	2
AUAAAU	1043	5	373.23	34.67	6	1529.15	12.43	4078
AAAAUA	1019	6	363.16	34.41	7	848.76	5.84	420
UAAAAU	1017	7	373.23	33.32	9	780.18	8.48	211
AUUAAA	1013	I	373.23	33.12	10	385.85	31.93	3
AUAAAG	972	9	184.27	58.03	4	593.90	15.51	34
UAAUAA	922	10	373.23	28.41	13	1233.24	–8.86	4034
UAAAAA	922	11	363.16	29.32	12	922.67	9.79	155
UUAAAA	863	12	373.23	25.35	15	374.81	25.21	4
CAAUAA	847	13	185.59	48.55	5	613.24	9.44	167
AAAAAA	841	14	353.37	25.94	14	496.38	15.47	36
UAAAUA	805	15	373.23	22.35	21	1143.73	–10.02	4068

Outline Update

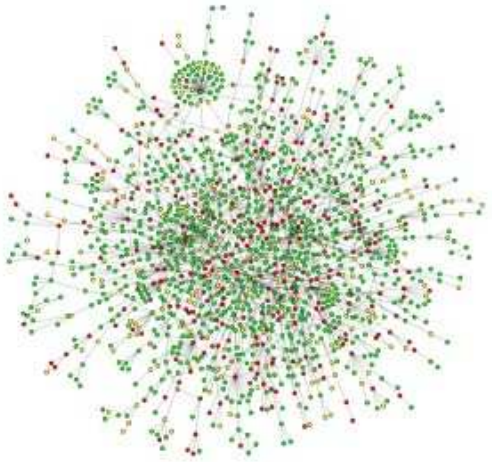
1. What is Information?
2. Information Transfer: Darwin Channel
3. Information Discovery in Massive Data: Pattern Matching
4. Information in Structures: Network Motifs
   - Biological Networks
   - Finding Biologically Significant Structures

Protein Interaction Networks

- Molecular Interaction Networks: Graph theoretical abstraction for the organization of the cell
- Protein-protein interactions (PPI Network)
  - Proteins signal to each other, form complexes to perform a particular function, transport each other in the cell...
  - It is possible to detect interacting proteins through high-throughput screening, small scale experiments, and *in silico* predictions



Undirected Graph Model

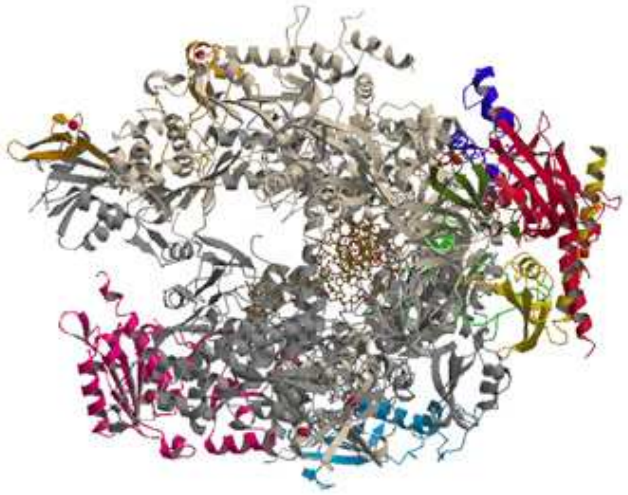


S. Cerevisiae PPI network hspace0.3in

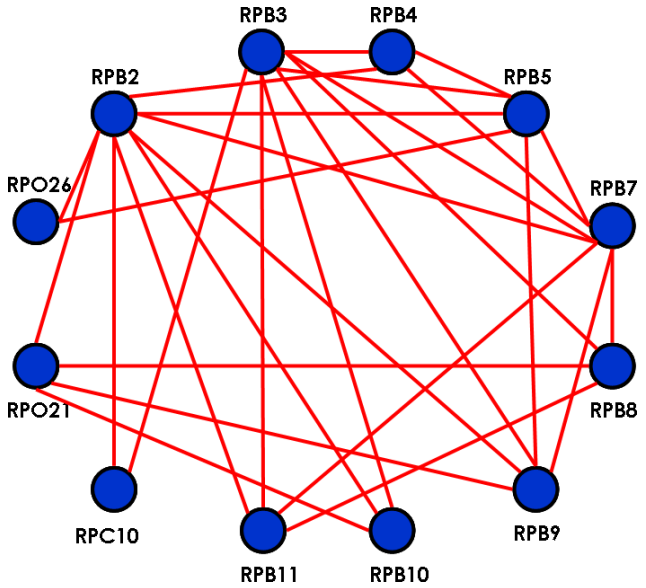
(Jeong et al., *Nature*, 2001)

Modularity in PPI Networks

- A functionally modular group of proteins (*e.g.* a protein complex) is likely to induce a dense subgraph
- Algorithmic approaches target identification of dense subgraphs
- An important problem: How do we define dense?
  - Statistical approach: What is significantly dense?



RNA Polymerase II Complex



Corresponding induced subgraph

Significance of Dense Subgraphs

- A subnet of r proteins is said to be ρ-dense if the number of interactions, $F(r)$, between these r proteins is $\geq \rho r^2$, that is,

$$F(r) \geq \rho r^2$$

- What is the expected size, R_ρ, of the largest ρ-dense subgraph in a random graph?
  - Any ρ-dense subgraph with larger size is statistically significant!
  - Maximum clique is a special case of this problem ($\rho = 1$)

- $G(n, p)$ model
  - n proteins, each interaction occurs with probability p
  - Simple enough to facilitate rigorous analysis

- Piecewise $G(n, p)$ model
  - Captures the basic characteristics of PPI networks

- Power-law model

Largest Dense Subgraph on $G(n,p)$

Theorem 6. *If G is a random graph with n nodes, where every edge exists with probability p, then*

$$\lim_{n\to\infty} \frac{R_\rho}{\log n} = \frac{1}{H_p(\rho)} \qquad (pr.),$$

where

$$H_p(\rho) = \rho \log \frac{\rho}{p} + (1-\rho) \log \frac{1-\rho}{1-p}$$

denotes divergence. More precisely,

$$P(R_\rho \geq r_0) \leq O\left(\frac{\log n}{n^{1/H_p(\rho)}}\right),$$

where

$$r_0 = \frac{\log n - \log\log n + \log H_p(\rho)}{H_p(\rho)}$$

for large n.

Proof

- $X_{r,\rho}$: number of subgraphs of size r with density at least ρ
  - $X_{r,\rho} = |\{U \subseteq V(G) : |U| = r \wedge |F(U)| \geq \rho r^2\}|$
- Y_r : number of edges induced by a set U of r vertices
  - $\mathbf{E}[X_r] = \binom{n}{r} P(Y_r \geq \rho r^2)$
  - $P(Y_r \geq \rho r^2) \leq \binom{r^2}{\rho r^2}(r^2 - \rho r^2)p^{\rho r^2}(1-p)^{r^2-\rho r^2}$
- Upper bound: By the first moment method:

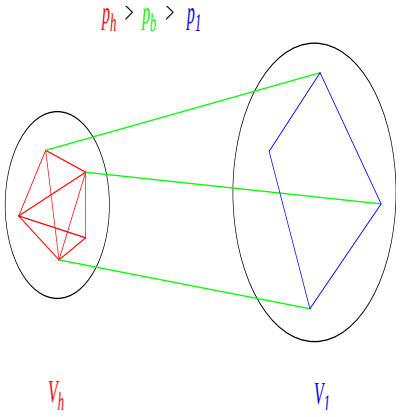
$$P(R_\rho \geq r) \leq P(X_{r,\rho} \geq 1) \leq \mathbf{E}[X_{r,\rho}]$$

  - Plug in Stirling's formula for appropriate regimes
- Lower bound: To use the second moment method, we have to account for dependencies in terms of nodes and existing edges
  - Use Stirling's formula, plug in continuous variables for range of dependencies

Piecewise $G(n,p)$ Model

- Few proteins with many interacting partners, many proteins with few interacting partners

Captures the basic characteristics of PPI networks, where $p_l < p_b < p_h$.



- $G(V,E)$, $V = V_h \cup V_l$ such that $n_h = |V_h| \ll |V_l| = n_l$

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h. \end{cases}$$

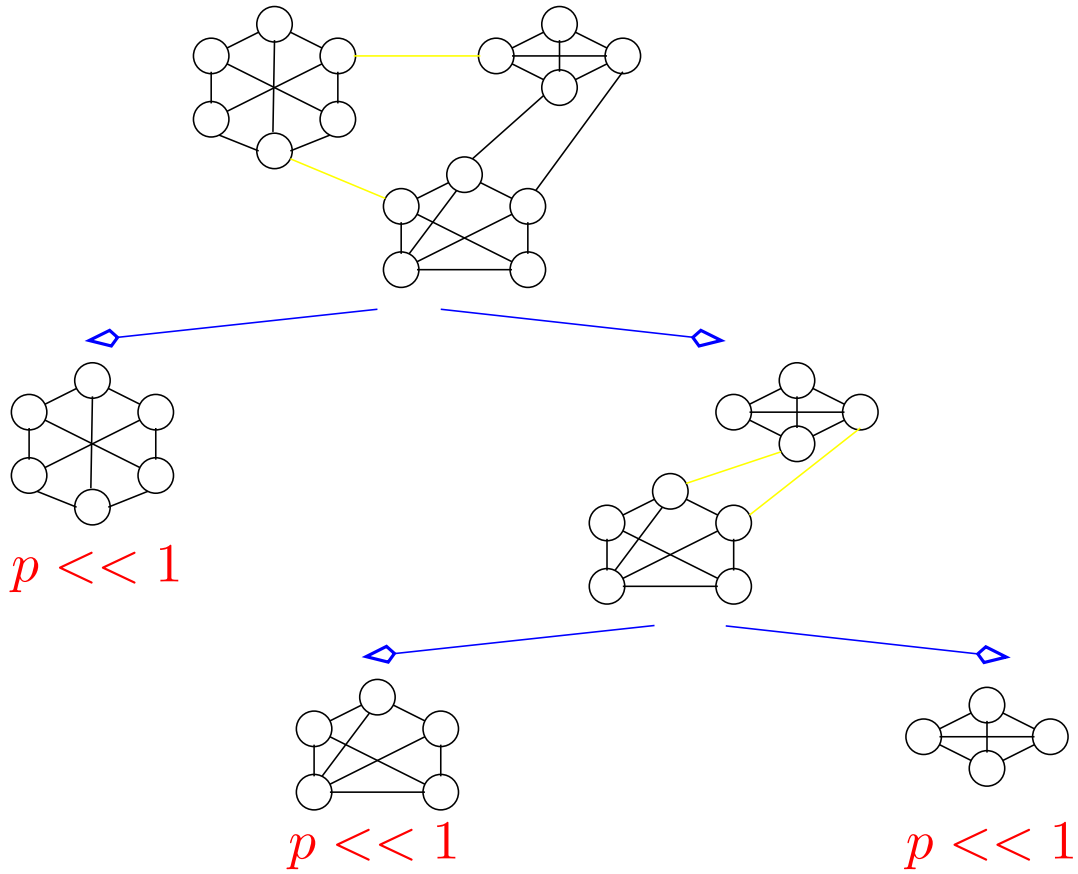
- If $n_h = O(1)$, then $P(R_\rho \geq r_1) \leq O\left(\frac{\log n}{n^{1/H_{p_l}(\rho)}}\right)$ where

$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log H_{p_l}(\rho) - \log e + 1}{H_{p_l}(\rho)}$$

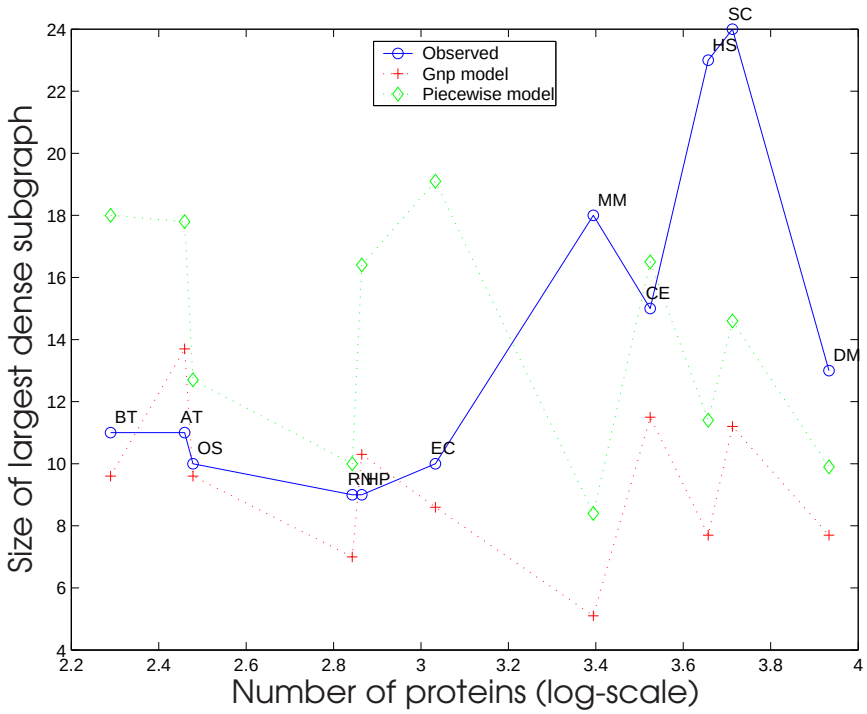
and $B = (p_b(1-p_l))/p_l + (1-p_b)$.

SIDES

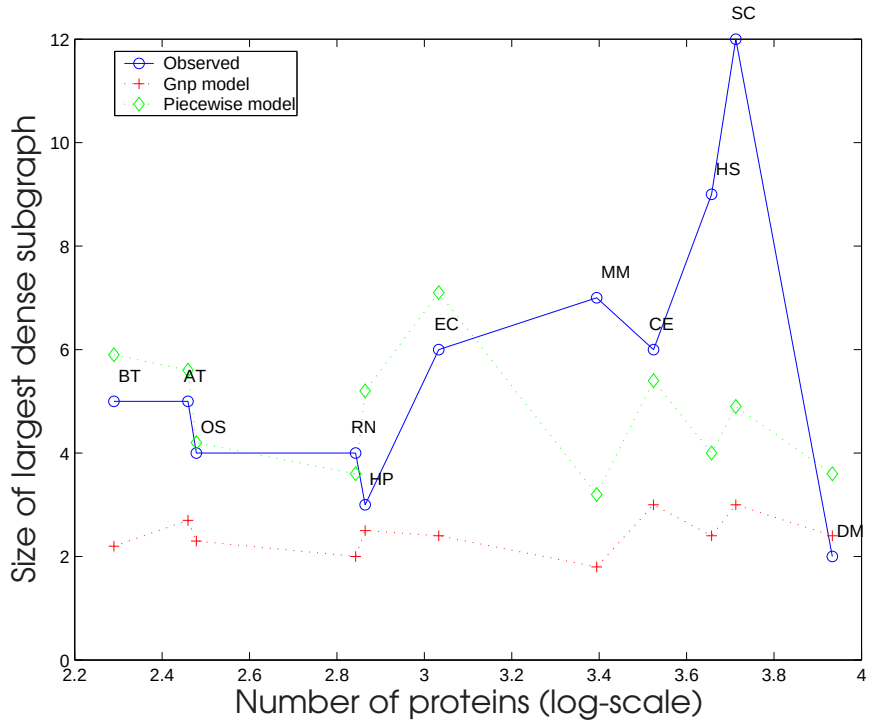
- An algorithm for identification of Significantly Dense Subgraphs (SIDES)
  - Based on Highly Connected Subgraphs algorithm (Hartuv & Shamir, 2000)
  - Recursive min-cut partitioning heuristic
  - We use statistical significance as stopping criterion



Behavior of Largest Dense Subgraph Across Species



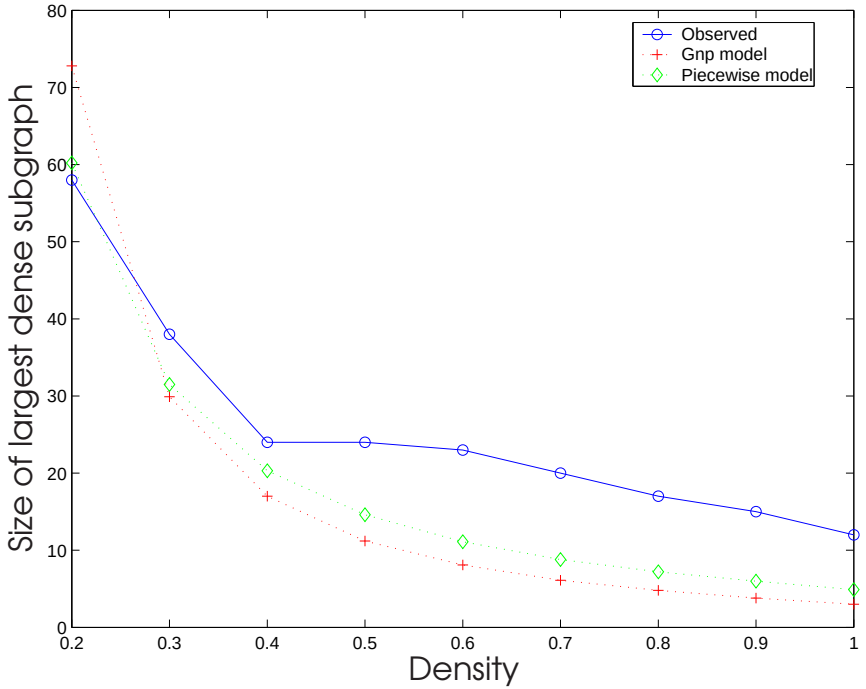
$\rho = 0.5$



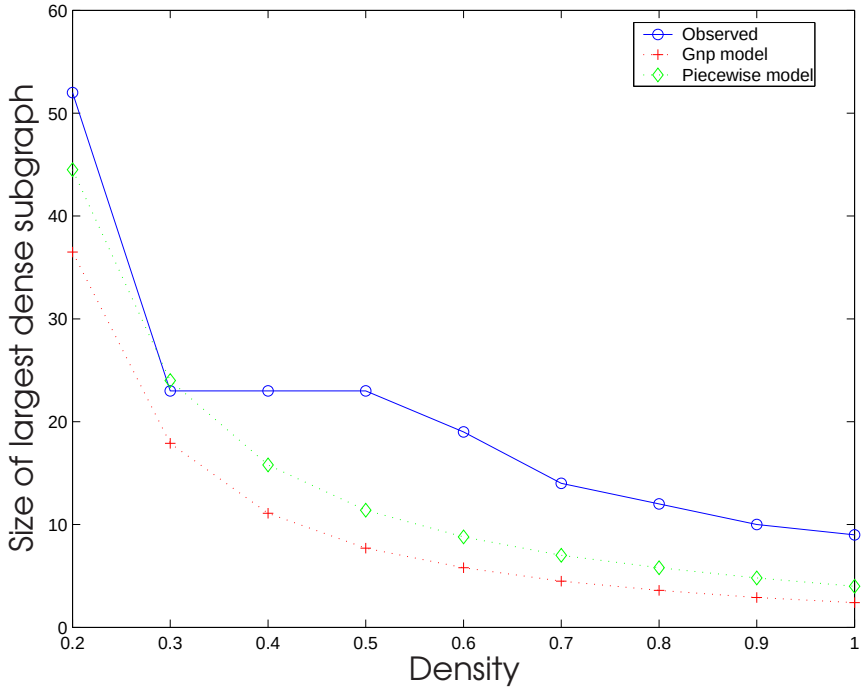
$\rho = 1.0$

Number of nodes vs. Size of largest dense subgraph
for PPI networks belonging to 9 Eukaryotic species

Behavior of Largest Dense Subgraph w.r.t Density



S. cerevisiae



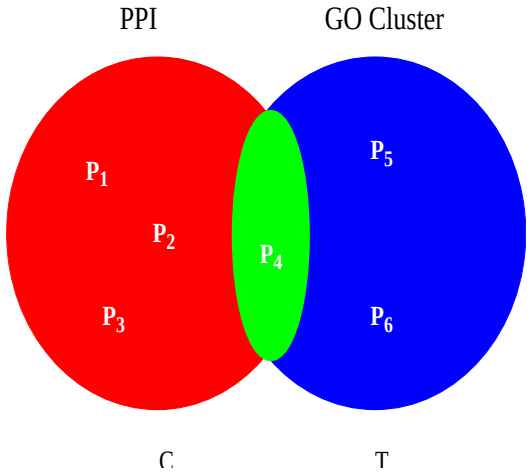
H. sapiens

Density threshold vs. Size of largest dense subgraph
for Yeast and Human PPI networks

Performance of SIDES

- Biological relevance of identified clusters is assessed with respect to Gene Ontology (GO)
  - Find GO Terms that are significantly enriched in each cluster
- Quality of the clusters
  - For GO Term t that is significantly enriched in cluster C, let T be the set of proteins associated with t

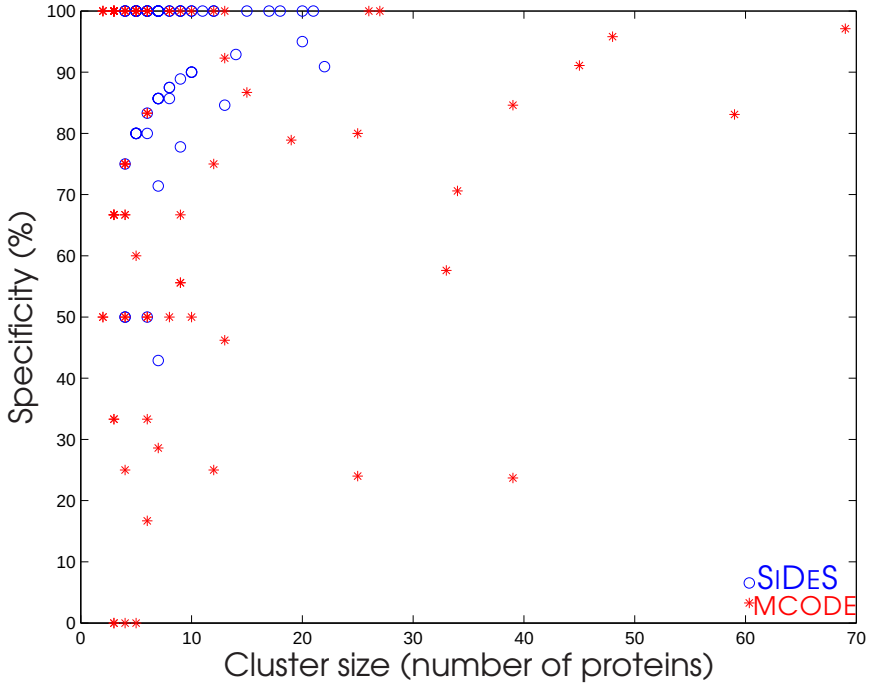
specificity = $100 \times |C \cap T|/|C|$
sensitivity = $100 \times |C \cap T|/|T|$

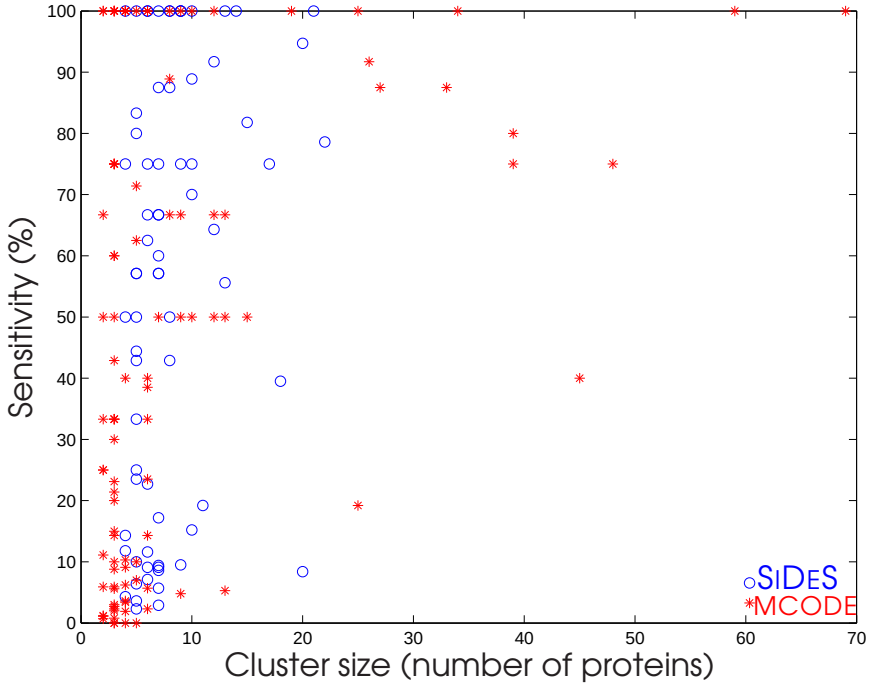


	SIDES			MCODE		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Specificity (%)	43.0	100.0	91.2	0.0	100.0	77.8
Sensitivity (%)	2.0	100.0	55.8	0.0	100.0	47.6

Comparison of SIDES with MCODE (Bader & Hogue, 2003) on yeast PPI network derived from DIP and BIND databases

Performance of SIDES





Correlation

SIDES: 0.22
MCODE: -0.02

SIDES: 0.27
MCODE: 0.36