A Gentle Introduction to Bioinformatics
The Invariant Perspective

Limsoon Wong
School of Computing &
Yong Loo Lin School of Medicine

What is an invariant?

- **Suppose you have a bag of x red beans and y green beans**
- **Repeat the following:**
  - Remove 2 beans
  - If both green, discard both
  - If both red, discard one, put back one
  - If one green and one red, discard red, put back green
- **If one bean is left behind, can you predict its colour?**

Shall we bet on the color of the bean that is left behind?

Bet on last bean being green

- **Suppose you have a bag of x red beans and y green beans**
- **Repeat the following:**
  - Remove 2 beans
  - If both green, discard both
  - If both red, discard one, put back one
  - If one green and one red, discard red, put back green
- **If one bean is left behind, can you predict its colour?**

- **When the parity of green beans is odd, it remains odd…**
- **Start with y=2n+1**
- **y=2n+1 → y=2n-1**
- **y=2n+1 → y=2n+1**
- **y=2n+1 → y=2n+1**
- **It must be green!**

Bet on last bean being red

- **Suppose you have a bag of x red beans and y green beans**
- **Repeat the following:**
  - Remove 2 beans
  - If both green, discard both
  - If both red, discard one, put back one
  - If one green and one red, discard red, put back green
- **If one bean is left behind, can you predict its colour?**

- **When the parity of green beans is even, it remains even…**
- **Start with y=2n**
- **y=2n → y=2n-2**
- **y=2n → y=2n**
- **y=2n → y=2n**
- **It must be red!**

Bet on color of the last bean … and win!

- Suppose you have a bag of x red beans and y green beans
- Repeat the following:
  - Remove 2 beans
  - If both green, discard both
  - If both red, discard one, put back one
  - If one green and one red, discard red, put back green
- If one bean is left behind, can you predict its colour?

- **If you start with odd # (even #) of green beans, there will always be an odd # (even #) of green beans in the bag**

⇒ **Parity of green beans is invariant**

⇒ **Bean left behind is green iff you start with odd # of green beans**

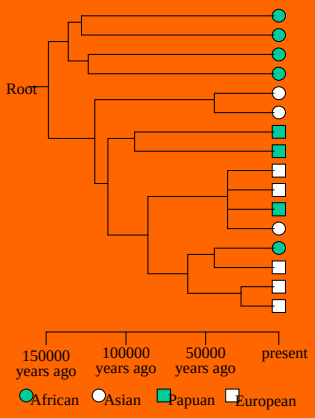
- **What have we just seen?**
- **Problem solving by logical reasoning on invariants**

A Gentle Introduction to Bioinformatics
The Invariant Perspective

Limsoon Wong
School of Computing &
Yong Loo Lin School of Medicine

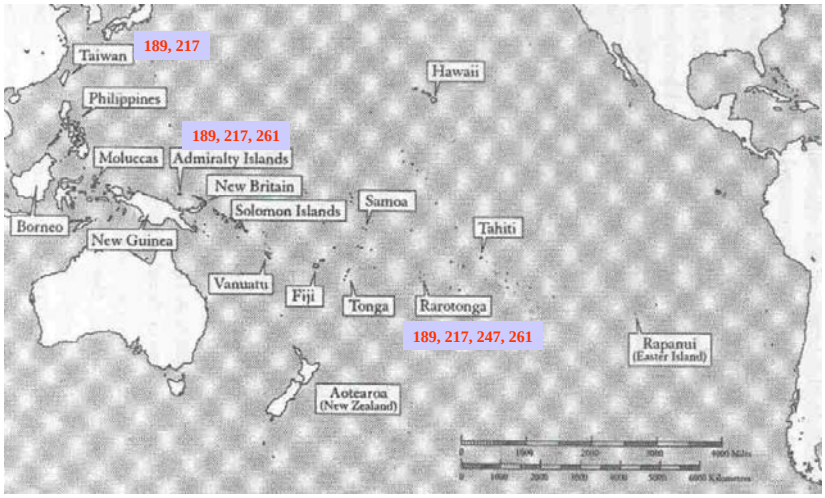
Plan

- **From Invariants to Origin of Species**
  - Where do Polynesians come from
- **Invariants in Evolution**
  - Finding Active Sites
- **From Invariants to Emerging Patterns**
  - Finding Key Mutation Sites
- **From Invariants to "Guilt by Association"**
  - Predicting Protein Functions
- **Invariants in Diseases**
  - Identifying ALL subtypes

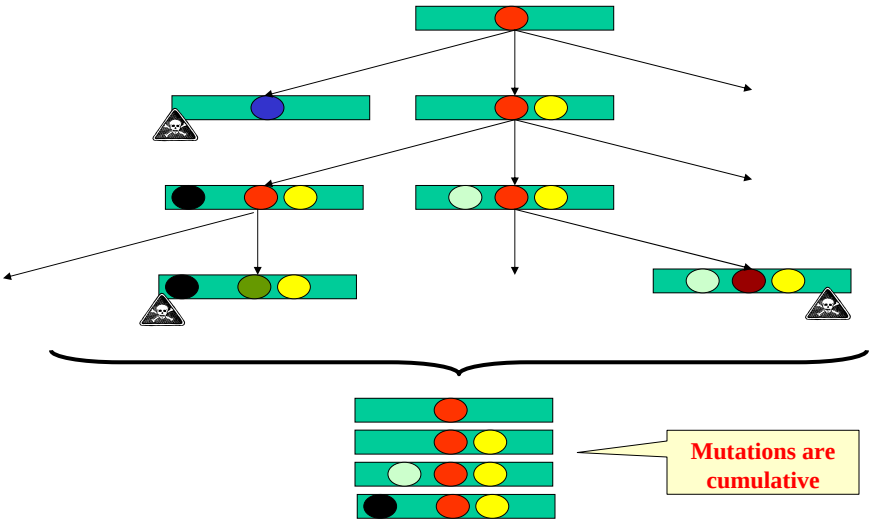


Where do Polynesians come from?

Do Polynesians come from Asia or America?



In the course of evolution…



What is the invariant?

- **Mitochondrial DNA accumulates 1 mutation about every 10,000 years**
- **Human history is not so long relative to this**

⇒ **When a nucleotide in mitochondrial DNA is mutated it stays mutated through future generations**

Origin of Polynesians

- Common mitochondrial control seq from Rarotonga have variants at positions 189, 217, 247, 261. Less common ones have 189, 217, 261
- Seq from Taiwan natives have variants 189, 217
- Seq from regions in betw have variants 189, 217, 261.

- More 189, 217 closer to Taiwan. More 189, 217, 261 closer to Rarotonga
- 247 not found in America

⇒ Polynesians came from Taiwan!

- Taiwan seq sometimes have extra mutations not found in other parts

⇒ These are mutations that happened since Polynesians left Taiwan!

Are Europeans descended purely from Cro Magnons? Purely from Neanderthals? Or mixed?

Neanderthal

Cro Magnon

Neanderthal vs Cro Magnon

- Based on palaeontology, Neanderthal & Cro Magnon last shared an ancestor 250,000 yrs ago
- Mitochondrial DNA accumulates 1 mutation per 10,000 yrs

⇒ If Europeans have mixed ancestry, the mitochondrial DNA betw 2 Europeans should have ~25 diff w/ high probability

- The number of diff betw Welsh is ~3, & at most 8.
- When compared w/ other Europeans, 14 diff at most

⇒ Ancestor either 100% Neanderthal or 100% Cro Magnon

- Mitochondrial DNA from Neanderthal have 26 diff from Europeans

⇒ Ancestor must be 100% Cro Magnon

The "Invariant" Perspective

- The invariant:

  When a nucleotide in mitochrondrial DNA is mutated it stays mutated through future generations

- The lesson learned:

Figure out origins of Polynesians and Europeans by logical reasoning on invariant

Invariants in Evolution

What is a domain

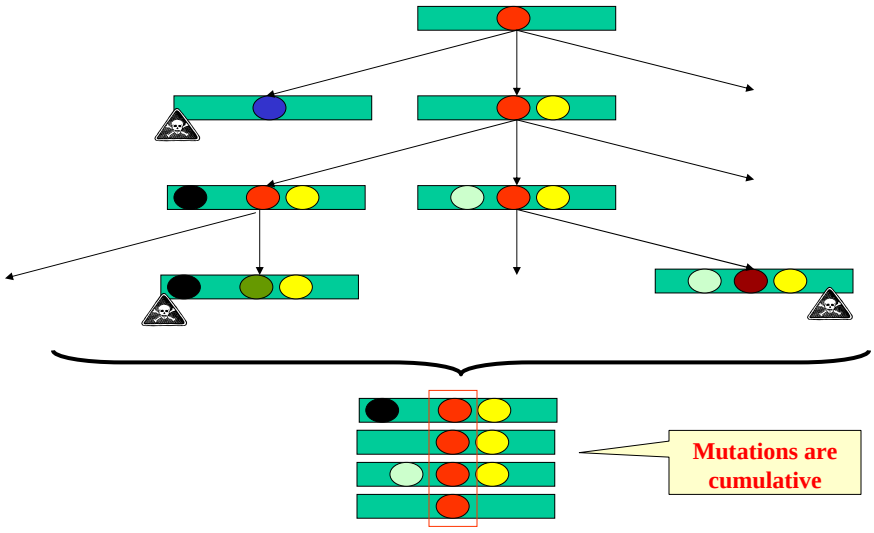
- **A domain is a component of a protein that is self-stabilizing and folds independently of the rest of the protein chain**
  - Not unique to protein products of one gene; can appear in a variety of proteins
  - Play key role in the biological function of proteins
  - Can be "swapped" by genetic engineering betw one protein and another to make chimeras
- **May be composed of one, more than one, or not any structural motifs (often corresponding to active sites)**

Discovering Domain and Active Sites

```
>gi|475902|emb|CAA83657.1| protein-tyrosine-phosphatase alpha
MDLWFFVLLLGSGLISVGATNVTTEPPTTVPTSTRIPTKAPTAAPDGGTTPRVSSLNVSSPMTTSAPASE
PPTTTATSISPNATTASLNASTPGTSVPTSAPVAISLPPSATPSALLTALPSTEAEMTERNVSATVTTQE
TSSASHNGNSDRRDETPIIAVMVALSSLLVIVFIIIVLYMLRFKKYKQAGSHSNSFRLPNGRTDDAEPQS
MPLLARSPSTNRKYPPLPVDKLEEEINRRIGDDNKLFREEFNALPACPIQATCEAASKEENKEKNRYVNI
LPYDHSRVHLTPVEGVPDSHYINTSFINSYQEKNKFIAAQGPKEETVNDFWRMIWEQNTATIVMVTNLKE
RKECKCAQYWPDQGCWTYGNIRVSVEDVTVLVDYTVRKFCIQQVGDVTNKKPQRLVTQFHFTSWPDFGVP
FTPIGMLKFLKKVKTCNPQYAGAIVVHCSAGVGRTGTFIVIDAMLDMMHAERKVDVYGFVSRIRAQRCQM
VQTDMQYVFIYQALLEHYLYGDTELEVTSLEIHLQKIYNKVPGTSSNGLEEEFKKLTSIKIQNDKMRTGN
LPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASFIDGYRRRTPTCQPRPVQHTIEDFWRMIWEWK
SCSIVMLTELEERGQEKCAQYWPSDGSVSYGDINVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFH
GWPEVGIPSDGKGMINIIAAVQKQQQQSGNHPMHCHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVK
SLRLQRPHMVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

- **How do we find the domain and associated active sites in the protein above?**

In the course of evolution…



Multiple Alignment of PTPs

```
gi|126467|      FHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|      YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906      FQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|      LHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|      FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715      FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
                 ..* *** ***      . *               ..****** ****... ** ..
```

- **Notice the PTPs agree with each other on some positions more than other positions**
- **These positions are more impt wrt PTPs**
- **Else they wouldn't be conserved by evolution**

⇒ **They are candidate active sites**

The "Invariant" Perspective

- **The invariant:**

  When an amino acid is impt for function, it is under evolutionary pressure to be conserved

- **The lesson learned:**

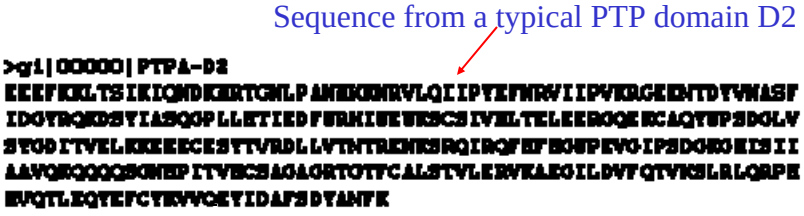
Figure out active sites by logical reasoning on invariant

A Twist in the Tale: From Invariant to Emerging Pattern

Identifying Key Mutation Sites

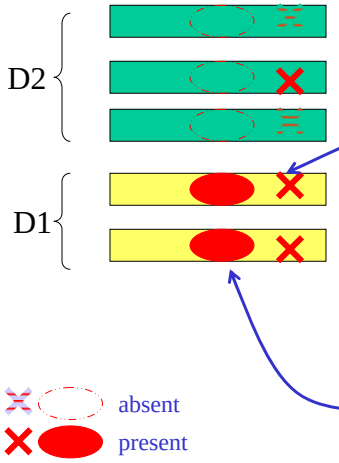
K.L.Lim et al., *JBC*, 273:28986--28993, 1998

Sequence from a typical PTP domain D2



- **Some PTPs have 2 PTP domains**
- **PTP domain D1 is has much more activity than PTP domain D2**
- **Why? And how do you figure that out?**

Invariant as Emerging Pattern



This site is consistently conserved in D1,
but is not consistently missing in D2
⇒ it is not an EP
⇒ not a likely cause of D2's loss of function

This site is consistently conserved in D1,
but is consistently missing in D2
⇒ it is an EP
⇒ possible cause of D2's loss of function

Emerging Patterns of PTP D1 vs D2

- **Collect example PTP D1 sequences**
- **Collect example PTP D2 sequences**
- **Make multiple alignment A1 of PTP D1**
- **Make multiple alignment A2 of PTP D2**
- **Are there positions conserved in A1 that are violated in A2?**
- **These are candidate mutations that cause PTP activity to weaken**
- **Confirm by wet experiments**

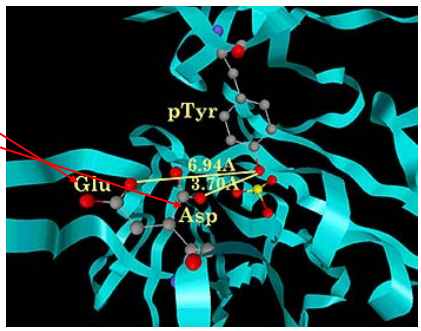
Key Mutation Site: PTP D1 vs D2

```
                    ?   !  ?         ?                  ?             ? ??
gi|00000|P  D2  QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|      QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|462550|      QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSML
gi|2499751      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
gi|1709906  D1  QFQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAML
gi|126471|      QLHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMM
gi|548626|      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|131570|      QFHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
                * ..  **. *.*      .                 . ****** ****..  .  ..
```

- **Positions marked by "!" and "?" are likely places responsible for reduced PTP activity**
  - All PTP D1 agree on them
  - All PTP D2 disagree on them

Key Mutation Site: PTP D1 vs D2

```
                    ?   !  ?
gi|00000|P  D2  QFHFHGWPEVGIPSDGK
gi|126467|      QFHFTSWPDFGVRFTPI
gi|2499753      QFHFTGWPDHGVPYHAT
gi|462550|      QYHYTQWPDMGVPEYAL
gi|2499751      QFHFTSWPDHGVPDTTD
gi|1709906  D1  QFQFTAWPDHGVPEHPT
gi|126471|      QLHFTSWPDFGVPFTPI
gi|548626|      QFHFTGWPDHGVPYHAT
gi|131570|      QFHFTGWPDHGVPYHAT
gi|2144715      QFHFTSWPDHGVPDTTD
                * ..  **. *.*
```



- **Positions marked by "!" are even more likely as 3D modeling predicts they induce large distortion to structure**

Confirmation by Mutagenesis Expt

- **What wet experiments are needed to confirm the prediction?**
  - Mutate E → D in D2 and see if there is gain in PTP activity
  - Mutate D → E in D1 and see if there is loss in PTP activity

The "Invariant" Perspective

- **The invariant:**

  When an amino acid is impt for function, it is under evolutionary pressure to be conserved

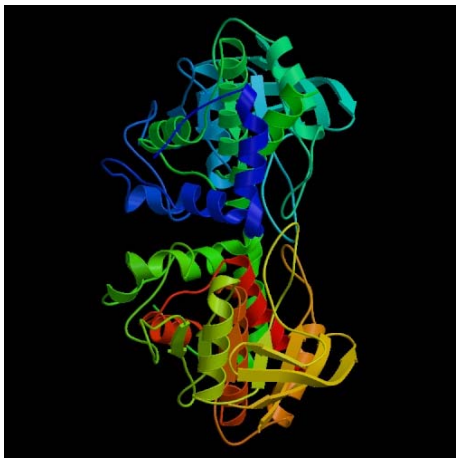
- **The lesson learned:**

Figure out key mutation sites by logical reasoning on invariant

From Invariant to Guilt by Association

A protein is a ...

- **A protein is a large complex molecule made up of one or more chains of amino acids**
- **Protein performs a wide variety of activities in the cell**



Function Assignment to Protein Sequence

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

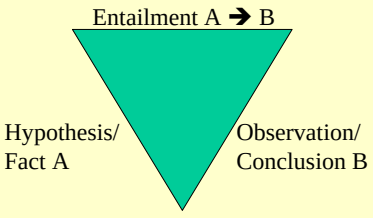
- **How do we attempt to assign a function to a new protein sequence?**

Invariant and Abductive Reasoning

- **Function is determined by 3D struct of protein & environment protein is in**
- **Constraints imposed by 3D struct & environment give rise to "invariant" properties observed in proteins having the ancestor with that function**

⇒ **Abductive reasoning**

- If those invariant properties are seen in a protein, then the protein is homolog of this protein



⇒ **"Guilt by association"**

Guilt-by-Association

- **Compare the target sequence T with sequences S_1, …, S_n of known function in a database**
- **Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T**
- **Then assign to T the same function as these homologs**
- **Finally, confirm with suitable wet experiments**

Sequence Alignment: Poor Example

- **Poor seq alignment shows few matched positions**

⇒ **The two proteins are not likely to be homologous**

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                      60        70        80        90       100
Amicyanin          MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                                :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFTYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                     70        80        90       100       110       120
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Sequence Alignment: Good Example

- **Good alignment usually has clusters of extensive matched positions**

⇒ **The two proteins are likely to be homologous**

```
>gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|   unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI  WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between
Amicyanin and unknown M. loti protein

Homologs obtained by BLAST

Sequences producing significant alignments:	Score (bits)	E Value
gi|14193729|gb|AAK56109.1|AF332081_1 protein tyrosin phosph...	623 L	e-177
gi|126467|sp|P18433|PTRA_HUMAN Protein-tyrosine phosphatase...	620 L	e-177
gi|4506303|ref|NP_002827.1| protein tyrosine phosphatase, r...	624 L	e-176
gi|227294|prf||1701300A protein Tyr phosphatase	620	e-176
gi|18450369|ref|NP_543030.1| protein tyrosine phosphatase, ...	624 L	e-176
gi|32067|emb|CAA37447.1| tyrosine phosphatase precursor [Ho...	619 L	e-176
gi|285113|pir||JC1285 protein-tyrosine-phosphatase (EC 3.1....	619	e-176
gi|6981446|ref|NP_036895.1| protein tyrosine phosphatase, r...	619 L	e-176
gi|2098414|pdb|1YFO|A Chain A, Receptor Protein Tyrosine Ph...	61 S	e-174
gi|32313|emb|CAA38662.1| protein-tyrosine phosphatase [Homo...	61 L	e-174
gi|450583|gb|AAB04150.1| protein tyrosine phosphatase >gi|4...	605	e-172
gi|6679557|ref|NP_033006.1| protein tyrosine phosphatase, r...	60 L	e-172
gi|483922|gb|AAA17990.1| protein tyrosine phosphatase alpha	599	e-170

- **Thus our example sequence could be a protein tyrosine phosphatase α (PTPα)**

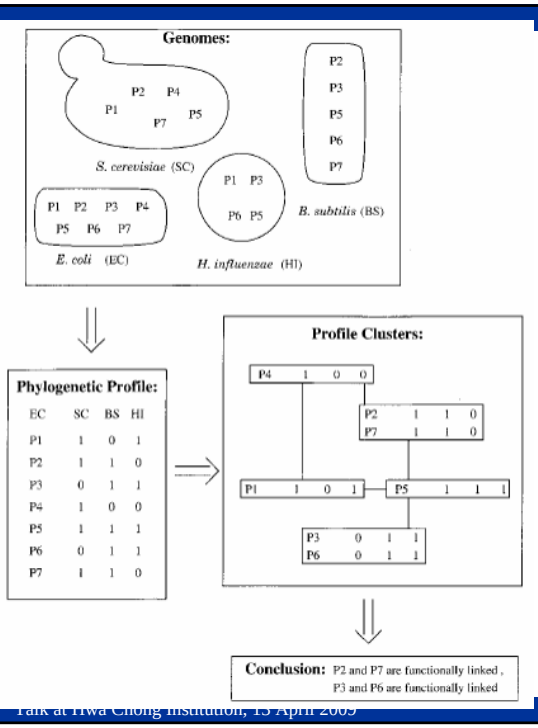
What if there is no sequence homology?

Guilt by association of other invariants of evolution!

Phylogenetic Profiles as Invariant

- **A protein is not alone when performing its biological function**

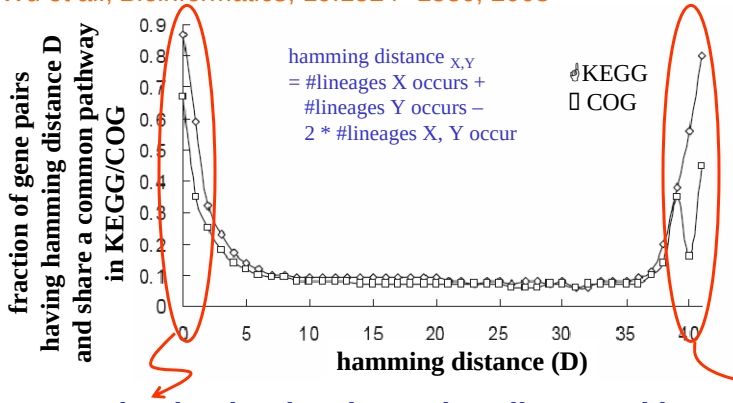
⇒ **Gene (and hence proteins) with identical patterns of occurrence across phyla tend to function together**



Phylogenetic Profiling: How It Works

Phylogenetic Profiling: Evidence

Wu et al., *Bioinformatics*, 19:1524--1530, 2003



- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

Exercise: Why do proteins having high hamming distance also have this behaviour?

The "Invariant" Perspective

- **The invariants:**

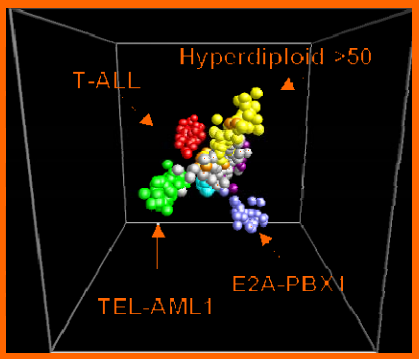
  When an amino acid is impt for function, it is under evolutionary pressure to be conserved

  When a partner is impt for function, it is under evolutionary pressure to be conserved

- **The lesson learned:**

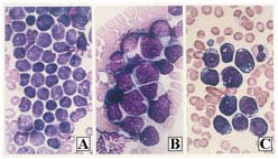
Figure out protein function by abductive reasoning on invariant

Invariants in Diseases



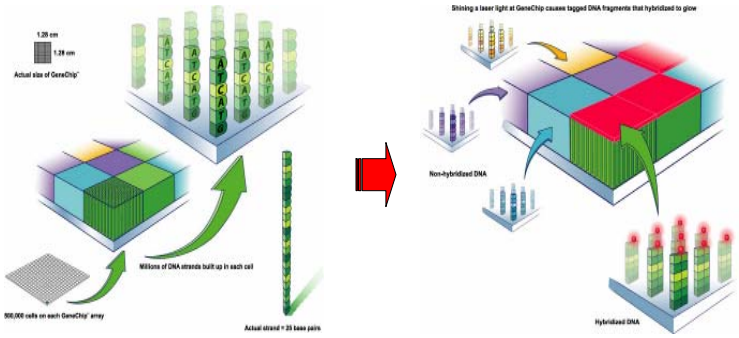
Childhood Acute Lymphoblastic Leukemia

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse
- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics

Massive Gene Expression Profiling



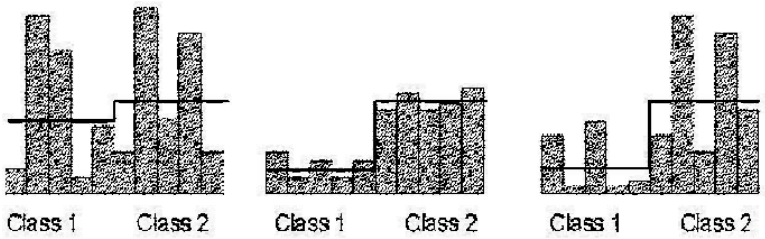
Subtype Diagnosis by Gene Expression

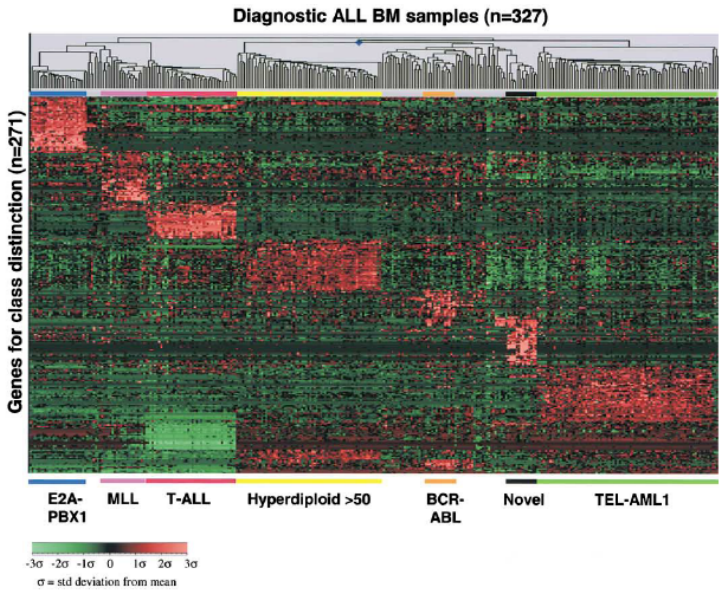
- **Gene expression data collection**
- **Gene selection by $\chi 2$**
- **Classifier training**
- **Apply classifier for diagnosis of future cases**

Signal Selection Basic Idea

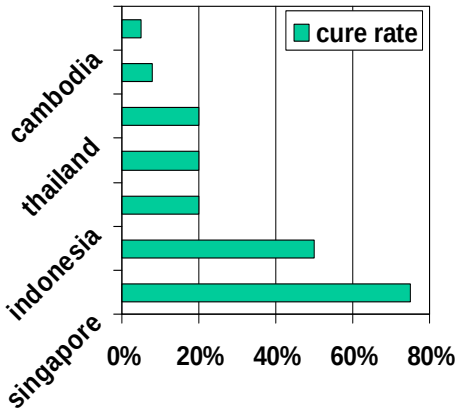
- **Choose a signal w/ low intra-class distance**
- **Choose a signal w/ high inter-class distance**

$\Rightarrow$ **Invariants which are emerging patterns**





Childhood ALL Cure Rates



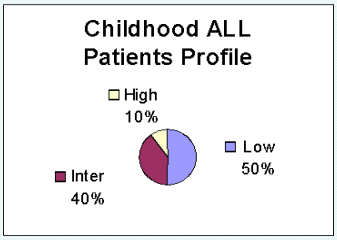
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

⇒ **Not available in less advanced ASEAN countries**

Childhood ALL Treatment Cost

- **Treatment for childhood ALL over 2 yrs**
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **Cost for side-effects: Unquantified**

Current Situation (2000 new cases/yr in ASEAN)



- **Intermediate intensity conventionally applied in less advanced ASEAN countries**

- **Over intensive for 50% of patients, thus more side effects**
- **Under intensive for 10% of patients, thus more relapse**
- **US$120m (US$60k * 2000) for intermediate intensity tx**
- **US$30m (US$150k * 2000 * 10%) for relapse tx**
- **Total US$150m/yr plus un-quantified costs for dealing with side effects**

Why not high/low intensity to everyone?

- **High-intensity Tx**
  - Over intensive for 90% of patients, thus a lot more side effects
  - US$144m (US$72k * 2000) for high-intensity tx

⇒ **Total US$144m/yr plus un-quantified costs for dealing with side effects**

- **Low-intensity Tx**
  - Under intensive for 50% of patients, thus a lot more relapse
  - US$72m (US$36k * 2000) for low-intensity tx
  - US$150m (US$150k * 2000 * 50%) for relapse tx

⇒ **Total US$222m/yr**

Exploit Invariant Gene Expr Profiles

- **Low intensity applied to 50% of patients**
- **Intermediate intensity to 40% of patients**
- **High intensity to 10% of patients**

⇒ **Reduced side effects**
⇒ **Reduced relapse**
⇒ **75-80% cure rates**

- **US$36m (US$36k * 2000 * 50%) for low intensity**
- **US$48m (US$60k * 2000 * 40%) for intermediate intensity**
- **US$14.4m (US$72k * 2000 * 10%) for high intensity**

- **Total US$98.4m/yr**

⇒ **Save US$51.6m/yr**

What have we learned?

What have we learned?

- **Paradigms**
  - Invariants
  - Emerging patterns
  - "Guilt by association"
- **Applications**
  - Active sites and key mutations
  - Origin of species
  - Protein functions
  - Disease diagnosis
- **Techniques**
  - Sequence comparison
  - Multiple alignment
  - Machine learning
  - Signal processing
- **Miscellaneous**
  - Microarrays
  - Economic of bioinformatics

Suggested Readings

- Limsoon Wong, *The Practical Bioinformatician*, World Scientific, 2004. Chapters 1, 3, 4, 14.
- K.L.Lim et al. “Interconversion of kinetic identities of the tandem catalytic domains of receptor-like protein tyrosine phosphatase PTP-alpha by two point mutations is synergist and substrate dependent”, *JBC*, 273:28986--28993, 1998
- J. Wu et al. “Identification of functional links between genes using phylogenetic profiles”, *Bioinformatics*, 19:1524--1530, 2003
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote homologies. *JCB*, 7(1-2):95—11, 2000
- B. Sykes. *The seven daughters of Eve*, Gorgi Books, 2002