Contents

Chapter 1

Evolutionary Biology

1.1 Concepts of Evolutionary Biology

Genotype-phenotype-fitness Paradigm: genetic variations within a population or across species lead to variations of phenotypes, and the phenotypic variations are subject to the force of natural selection. These variations (of genotype, phenotype and survival) can be used to understand the relationship among the three.

- Genotypic variations: could be created/influence from many sources: mutation, random drift, migration, etc.

- Phenotypic variations: both genetic and non-genetic factors. The critical issue is how a phenotypic trait is related to the underlying genotypic changes, e.g. many genotypic changes may have no phenotypic manifestation because of feedback controls.

- Fitness: affected by phenotypic traits, interaction with environments, etc. The relation between traits and fitness is also a critical issue: the fitness of a trait may depend on other traits, a trait may be good in some environments, but not the others, etc.

Problems of evolution: the goal is to explain observations of biological systems (novel functions, design features, variations) in the evolutionary framework:

- Intra- and inter-species variations and constraints (lack of variations):
  - What are the main constraints of evolution of a system? Ex. interacting partners should co-evolve.
  - Sources of variations and conservation: what sources are more important for genetic and phenotypic variations? Two main hypothesis: neutral theory - most variations from neutral mutations, and adaptationist theory - most variations are adaptations for micro-niches. What are the causes of unexpected level of conservation?

- Understanding the robustness and fragility of system:
  - Why some changes are tolerant and some others not?
  - What features may facilitate evolution, i.e. increase the mutational robustness? Ex. negative feedbacks that stabilize the output of a system.

- Understand how the fitness depends on genotypes and phenotypes: through molecular understanding and through exploring the variations of genotypes, phenotypes and fitness (different traits are selected differently within or across populations).
  - What genotypes/phenotypes are preferred under a given environment?

  - How do the organisms adapt to the changing environments (they need to do well in the current environment, but also do well when the environment changes)?

- Understanding design features of a system from an evolutionary perspective (design is a consequence of evolution):

  - There may be many ways of implementing the same function, why a particular design is chosen? Does evolution optimize anything in the design? To reformulate, if repeat evolution twice, will the same design emerges (chance vs necessity)?
  - What is the main determinant of a design feature: natural selection or physiological/developmental constraint?

- Adaptation and evolution of novel functions/phenotypes:

  - Selection of novel functions: What is the selective force of new functions? Is adaptation driven by environmental changes (including species interactions), or by internal "inventions"?
  - Evolutionary path of novel functions: What is the genetic basis of adaptation: proteins, gene expression control (trans- and cis- changes) or gene networks? What is the evolutionary path: each step must be positively selected or at least not harmful? What are the possible fitness barriers and how are they overcome? This is particularly important when the new functions conflict with the old functions, e.g. evolution of mimicry - good for avoiding predators but not good for mating.
  - Evolution of coordinated changes: e.g. when changing body size, not just bones, but also internal organs will also change proportionally. How are these coordinated changes obtained? What is the genetic architecture, a few master regulators, or many regulators controlling each of the organs? Ex. dog body size evolution (through artificial selection).

- The major evolutionary events: the origin of life, the evolution of sex, speciation, the evolution of altruism, etc.

Methods of evolutionary biology:

- The comparative analysis of genotypic and functional data, the ecological niches, etc. from different species and/or within populations: explain the comparative data from the knowledge of how the system works and the possible selection forces. Typically, starts from characterizing the changes or conservation of the system across species/individuals, then infer the underlying forces that constrain or change the system (e.g. new evolutionary challenges that need adaptation). The difficult part is that often the selection force is not clear.

- Cross-species data reveals history of evolution: e.g. ribosomal RNA sequences can be used to infer tree of life; the homologous relationship between photosystem in plant and the related system in cynobacterial suggests the common evolutionary origin.

- The fossil records and the climate/geographical data provide information on the physiology, development and environment of organisms in the past.

- Experimental method: test specific hypothesis, e.g. if a process involves an intermediate stage, then this stage may be experimentally reconstructed or some crucial assumption of this stage may be experimentally tested.

- Experimental evolution: evolve the organisms in experimental conditions.

Fundamental Principles of Evolution:: the interaction between basic design considerations (trade-offs, limits of physiology/development, etc.) and natural selection determines the structure, function and behavior of biological systems.

- Adaptation vs spandrel: generally a biological feature serves some purpose (adaptive), however, this is not always true: some may be a by-product of evolution, some may have a different cause. Ex. the adaptive value of dreaming is not clear; sneezing during bacterial infection: could be an adaptive host response, but could also be the result of manuipulation of the host by the pathogens to increase their spread.

- Optimization and trade-offs: Natural selection may optimize certain aspects of the organisms. On the other hand, any design almost always invovles trade-offs, e.g. size-mobility trade-off of most organisms, stability-function trade-off of proteins, specificity-sensitivity trade-off of signaling systems. Thus, to analyze a system: suppose some property is changed, analyze how other parts of the system is affected (trade-off), and the fitness consequence is the final results of all the changes; also the environment determines how trade-off is acted upon by natural selection. Ex. increasing size in insects will make oxygen supply difficult (surface-volume ratio is reduced), thus reducing mobility; in mammals, the trade-off is more favorable to size because of a different oxygen supply system (lung).

- Random tinkering: evolution can only act upon existing components. The consequences are: (1) evolution tends to reuse existing components, even for a different purpose; (2) the design may not always be optimal: e.g. the relative position of retina and optic nerve [Darwinian medicine]; (3) some features may not be possible to evolve, if a large change of the system is required.

- Evolution occurs in multiple levels of biology: changes at one level (e.g. DNA sequences) may not affect the next level (e.g. gene expression). A special case is compensatory changes at a lower level so that the output at the higher level is constant.

- Evolutionary strategies for interaction with other individuals or species: certain strategies will be chosen in the presence of species/individual interactions and these strategies have evolutionary consequences in the physiology and behavior of organisms. Arms race, symbiosis, cooperation, competition, etc.

- Mal-adaptation may occur: when the environment changes, some current features/behaviors may become mal-adaptive, e.g. many aspects of human biology were evolved not for modern life [Darwinian medicine, Survival of the Sickest].

- Dealing with changing environment: the the organisms should evolve to maximze average fitness. In a constantly changing environment, generalists are favored; in contrast, in a relatively constant environment, a specialist may be favored. Ex. Lack Vitorial, the diversity of cichild fish (relatively constant environments thus support many specialists).

- Variations: Mechanisms promoting genetic/phenotypic variations may be evolutionary advantageous: the mechanisms include increasing mutations/recombinations (e.g. VDJ recombination), increasing phenotypic variation (e.g. Hsp90 inactivation, stochastic gene expression). The benefits include: (1) environmental uncertainy: ensure that at least some will survive; (2) frequency-dependent selection: some rare alleles/phenotypes may be advantageous, esp. in co-evolution (e.g. between host and pathogen, the host system is evolve for fighting common pathogen alleles). Note that these explanations do not invoke group selection.

How did novel phenotypes evolve?

- A complex phenotype occurs through multiple stages/steps. At each stage, some trait/feature is evolved, which opens up possibilities to further "improve" the system. However, this trait/feature itself must be adaptive (or at least not detrimental) at the present condition. Example: bipedalism evolves before language, tool-making, etc., for other reasons: more energy efficient forager, etc. Also note that the intermediate states may not be observable in extant data.

- Co-option of existing modules for a different purpose: this includes the case where gene duplication creates the opportunity of divergence of functions. E.g. the ion channels important for nerve cells exist long before the evolution of nervous systems.

- Adaptation of complex systems: may be achieved through how different modules/units are linked to each other. E.g. cross-resistence to different stresses: linking the regulators (of one stress) and effector genes (of another stress).

- Evolvability: certain systems are highly evolvable. The key features of the evolvability may include: modular design (e.g. eye), master regulators controlling entire modules , including pleiotropic effects of single genes (e.g. Pax-6 for eye evolution). Ex. developmental system: the nervous system is in the dorsal side in vertebrates but ventral side in arthopods, may involve the change of a few signaling pathways (Dgg signaling in Drosophila and BMP4 signaling in vertebrates).

- Convergent evolution: different evolutionary paths may converge to the same solution of the same problem. Thus the similarity of the system reveals convergent design, instead of the common ancestral relationship.

How did a system evolve/change?

- Environmental changes: e.g. increase of glucose transporters in response to glucose limitation.

- Coordinating with changes in other system: e.g. nitrogen metabolism needs to be coordinated with the change of carbon metabolism.

- Change of some components may open new evolutionary paths: e.g. a stronger transcriptional activator becomes available (from other selection forces), it may replace some current activators (Tbf1/Cbf1 to Rap1 transition may fall in this category).

- Neutral changes could become adaptive when the environment changes; or a trait that is adapted for one purpose may be used for another purpose later at a different environment (ex-adaptation). In general, the genetic variations in a population (exist before any selection) provide sources for adaptation. Ex. type III secretion system was later used for flagellar.

Organizational features of biological systems: consequences of evolutionary principles discussed above:

- Non-optimal design: the best solutions may not always be reachable due to constraints of the mutational process/evolutionary path, which favors the increase of entropy (more mutationally likely). The result is a balance of the two forces: the equalibrium probability of a state depends on both its fitness, and the probability of mutation that leads to this state (mutation-selection balance in the case of large population). Similarly, the chosen design may not be parsimonious.

- Mutational robustness: the changes of structure from mutations may be tolerated because of inherent feedbacks, or the changes may compensate each other (not strictly neutral, but the consequence is that the phenotype does not change), or the changes of traits (within certain ranges) do not have fitness consequences. The robustness of systems may be enhanced by "evolutionary capacitor" such as heat shock proteins. The features that enable mutational robustness may be evolved as an adaptation to changing environments (s.t. the system functions well in many conditions) [Wagner, Robustness and evolvability].

- Multiplicity and redundancy: a common feature of many systems. The redudancy may be relative, e.g. in the case of isozymes, they may appear redundant in some conditions, but not in other conditions. The redundancy may be evolved as a consequence of duplication-specialization: the multiple specialists may still be able to compensate each other under some conditions.

- Modularity: this may be a consequence of reusing existing elements/components by evolution, and/or natural selection of functional subsystems. Ex. signaling networks: selection may add additional proteins to an existing signaling pathway (e.g. to make it respond to new signals); and the pathway may be used as a module in a larger network through cross-linking related pathways.

1.2 Robustness and Evolvability in Living Systems (Wagner)

1. Robustness and evolvability: overview

   Reference: [Wagner, Robustness and evolvability in living systems, Chap. 1, 2005]

   Biological variations:

   - Variations across differnet individuals and under different environments is a defining feature of biological systems. Most engineering systems would not have this design challenge (mass produciton so products tend to be the same, structurally more stable, etc.).
   - Genetic variations: relatively uniform across different genes.
   - Environmental variations: the changes of the system in response to the environmental changes can be highly correlated (e.g. all growth-related genes may affected at the same time).
   - Stochasticity: even if everything is equal, the system may have internal variations due to stochasiticity of chemical reactions, in particular, gene transcription.

   Canalization and robustness:

   - Canalization and robustness: different genotypes, under different environments/perturbations, lead to the same phenotype. In the context of development, it is often called canalization; in the context of gene networks, often called robustness. Canalization or robustness to variations, is a fundamental property of biological systems.
   - Robustness from a design perspective (genome as coding/instructions): the function of a system can be implemented by many different equivalent designs/solutions.
   - Robustness from a dynamic system perspective: the system has a few steady states, with potentially large basin of attraction.

   Mechanisms of robustness:

   - Redundancy: e.g. gene duplication in metabolic neworks and TRN, multiple binding sites in a CRS, two kidneys in a human body, and so on.
   - Compensation: one part of the system may compensate another part. In general, when we have two subsystems working in tandem, A followed by B. Then when the function of A is compromised, the input to B is reduced, and this is compensated by the increased function of B. Examples:
     - Metabolic networks: the deficiency of one enzyme/pathway (e.g. synthesis of an amino acid) can be compensated by a different pathway (e.g. uptake from the environment or synthesis from a different route).
     - Defense of human body: when skin is compromised, the immune system could take up and help defend the body.
   - Control of behavior: this is particularly evident when the system has a all-or-none behavior (e.g. achieved by cooperative interaction of molecules). Other examples of control include cells committed to a certain fate (differentiation) and cannot be reversed.
   - The role of feedbacks: provide a mechanism of sensing the problems in other parts of the system so that compensation by redundant components or a different part can take place.

   Evolutionary origin of robustness:

   - Functional requirements: e.g. the all-or-one behavior of cellular response to signals also makes it robust to mutations. Implementing a function often requires multiple subsystems, which could compensate each other.

- Selection for robustness to genetic changes (no need of invoking group selection): a robust system could tolerate many different mutations at different genes(alternative alleles), thus a mutation that makes the system more robust will have higher average fitness (some form of genetic interaction). Note: this is similar to Wagner's argument of mutational robustness (alternative designs are in the neutral space, and mutationally connected).
- Selection for robustness to nongenetic changes: phenotypic stability (e.g. developmental canalization) is advantageous in a fluctuating environment.
- Evolutionary consequence/history: duplication followed by specialization is common in evolution, and this gives some redundancy, hence robustness to the system. A robust system may be more evolutionary accessible (the concept of neutral space).
- Relation to phenotypic diversity/plasticity: in some cases, diversity may be preferred over canalization, e.g. stochastic gene expression may be advantegous in a fluctuating environment as a hedging strategy.

Evolutionary consequence of canalization:

- Evolutionary capacitor: robustness of a system allows it to hide many genetic and epigenetic variations (all phenotypically equivalent). However, when the conditions become extreme, canalization may be broken, and the phenotypic diversity may be resulted, which are subject to selection. This leads to the possiblity of innovations (evolability of the sytem).
- Neutral space: a related explanation of evolutionary innovation as a consequence of canalization. Because many solutions are equivalent, it is easy for the sytem to change to a related solution via mutations, and thus explore more solutions in the space (which can have different phenotypes when the conditions are different).
- Genetic assimilation: an initial environment-induced phenotype may be selected and encoded in the later generations even without stimulus. This may be explained by canalization (new phenotypes at extreme conditions), however, it may also be explained by quantitive genetic models.

Principles of canalization/robustness [Wagner]:

- Foundation: Most problems a living system has to solve have many equivalent solutions (Chap 13).
- Genetic origin of robustness: Evolution tends to converge to mutationally robust systems: first, they are more likely to find (Chap 13); second, mutational robustness can be increased by incremental evolution (Chap 16).
- Non-genetic origin of robustness: non-genetic change can drive incremental evolution of mutational robustness (Chap 17).
- Mechanism of robustness: redundancy of a system's parts and distributed robustness (Chap 15).
- Robustness is a key to evolutionary innovation (Chap 14).

2. Robust biological systems [Wagner, Chapter 7-10, 2005]

   Regulatory DNA sequences:

   - Regulatory sequences are robust: many changes in regulatory sequences, still preserve the expression pattern.
   - eve strip 2 enhancer:
     - Chimeri enhancer from Dmel and Dpse produces expression pattern different from wild type (slight expansion). Furthmore, Bcd-3 is essential in Dmel, but not in Dpse. These results suggest that the mutations change parts of the enhancers, but the compensatory changes in the two parts maintain expression.

  - Overall, still more constrained than introns.
- endo-16 enhancer of sea urchin:
  - Highly conserved expression pattern, but very different regulatory sequences. In one species, the TFBSs cannot be identified.
  - Placing one enhancer in another species: change of expression pattern, suggesting that TF expression/activitity or even the relevant TFs have changed.
  - Regulatory sequences have almost neutral rate of changes.
- unc-119 enhancer of worm: also highly divergent sequences with conserved expression.

Metabolic pathways:

- Background: consider a linear pathway: $\mathrm{S} \longrightarrow \mathrm{m}_1 \longrightarrow \mathrm{m}_2 ... \longrightarrow \mathrm{m}_{\text{n-1}} \longrightarrow \mathrm{P}$, and let E_i be the enzyme of the i-th step, and K_i be the total equilibrium constant of the first i reactions. The goal is to analyze how the flux through this pathway F, depends on the change of the activity of individual enzymes. Define the control coefficient of E_i as:

  $$C_i = \frac{\partial F}{\partial E_i} \tag{1.1}$$

  Thus large C_i means changing E_i has a large effect on the flux. The flux is related to the enzyme activities through:

  $$F = \frac{M}{1/E_1 + 1/E_2 + \cdots 1/E_n} \tag{1.2}$$

  where M is a constant determined by [S], [P] and K_n, and $E_i = K_{i-1}V_{m,i}/K_{m,i}$, where $V_{m,i}$ is the maximum velocity of i and $K_{m,i}$ is the M-M constant of i. From this, one could show that:

  $$C_i = \frac{1/E_i}{1/E_1 + 1/E_2 + \cdots 1/E_n} \tag{1.3}$$

  And $\sum_i E_i = 1$, where $0 \leq C_i \leq 1$. This relationship suggests a *distributed flux control* for linear pathways: the control coefficient of one enzyme is $1/n$ on average, a small value for long pathways.
- Robustness of linear pathways: because of distributed control, changing enzyme activity has a small effect on the flux, e.g. $n = 10$, changing enzyme activity by two-fold changes flux by less than 10%. In the evolutionary context, the flux of the pathway may fall in the "neutral zone", where changing enzyme activity has a small effect on fitness (which is assumed to be dependent on the flux). This is true if evolution has optimized enzyme activities to achieve the desired flux in the past, and then may cease to constrain the enzyme activities once it enters the "neutral zone".
- Evidence of robustness in metabolic pathways:
  - Dominance: for most enzymes, any single copy of enzyme is functional. This suggest that the function is insensitive to the dosage of individual enzymes.
  - Enzyme polymorphism: the polymorphism data can allow one to measure the fitness effect of mutations (mutation-selection balance), and the effect is small for the ADH enzyme in Drosophila.
  - Distribution of flux control coefficients in a pathway: the theory predicts that most enzymes have small control coefficients while a few have large values. This is confirmed by several pathways.
- Evolution of metabolic pathways: the enzyme activities may undergo random drift in the neutral space. Over time, it may be possible that once an important enzyme under selection may become neutral and vice versa.

- Remark:
  - Distributed flux control: the intuition is that changing activity of one enzyme (say, increase by two fold), leads initially a two-fold increase of flux through this enzyme, but it will be used to increase and maintain the higher level of all intermediate reactions at the new steady state, thus the overall flux increase is small.
  - An open question: whether changing in selection or neutral drift contributes to the pattern of moleular evolution (polymorphism).

Metabolic networks:

- Flux-balance analysis: suppose S represents the stoichoimetry matrix, $\mathbf{v}$ represents the flux in the network (one value per reaction) and $\mathbf{b}$ represents the external flux, then the flux must satisfy the flux balance equation (mass balance): $S\mathbf{v} = \mathbf{b}$. Thus all flux allowed would fall in the null space of the matrix S. The actual fluxes are often constrained by a number of other factors, e.g. some reactions are irreversible, the enzyme levels, etc. Additional considerations include: the different substrates/products (e.g. AAs and ATP) are balance in real biological systems to avoid waster. To determine the flux distribution in a given situation, it is assumed that certain objective function is maximized/minimized, commonly linear function:

  $$Z(\mathbf{v}) = \sum_i c_i v_i \tag{1.4}$$

  The objective function can be: the growth rate under a certain condition (e.g. carbon source, oxygen level).
- Robustness to elimination or reduction of enzymes: in E. coli, elimiate or reduce activities of enzymes in the central metabolism (including glycolysis, PPP, TCA cycle and respiration), and measure growth rates under different conditions:
  - Only 7 of the 48 reactions are essential, for the rest, most have small growth effects. One interesting case is the deletion of the first enzyme in PPP: little phenotypic effect through large metabolic changes. (1) Increasing flux through TCA cycle, to gain more NADH, and a large increase of flux through NADH $\longrightarrow$ NADPH. (2) Generate ribose-5-phosphate through other intermediates of glycolysis.
  - Even large reduction of essential reactions may have small fitness effects: e.g. reduction of more than 80% activities of the first three reactions of TCA leads to metabolic changes: (1) increase of PPP flux to obtain more NADPH; (2) reduction of flux through pyruvate kinase and TCA. As long as there is enough αKG (unless the flux through citrate synthase falls below 18%), the maximal growth can be maintained.
  - It is important to note that the fitness effect may depend on the species (e.g. less robust in H. influenzae) and the conditions.
- Robustness and elementary model analysis:
  - Elementary mode: a certain flux pattern through the network where only one path has flux. A large network may have a large number of elementary modes.
  - Elimination of enzymes reduces the number of elementary modes in the network. This number correlates with the robustness of the network.
- Evoultion of metabolic networks:
  - Why network topology is not a good predictor of evolutionary rates? It was suggested that hub proteins evolve more slowly. However, the terminal ones often are responsible for essential products that cannot be replaced through metabolic re-configuration.
  - An important determinant of the protein evolutionary rate of metabolic enzymes is the flux through this enzyme: larger flux are correlated with lower rates. Also note that gene duplications may lead to larger flux, thus also correlates with the rates.

Gene regulatory networks:

- Segment polarity network: the expression patterns of pair-rule genes, eve, ftz, odd, etc. creates the expression pattern of segment polarity genes: en (in anterior of each segment) and wg (posterior of each segment), and the segment polarity pattern is maintained (unlike pair-rule genes, whose patterns are transient) through signaling and transcriptional networks of the adjacent cells (which interact through ligand-receptor interactions).
- Robustness of segment polarity network: the expression patterns of en and wg are robust.
  - To parameters: random sampling of parameters, by chance, probability 0.9 that a parameter can produce a functioning network. Also for a given parameter setting, changing value of one parameter often can be tolerated.
  - To expression patterns of pair-rule genes: even if they do not have sharp boundaries.
  - To network topology: some topology are robust, but others are less robust.
- Network features important for robustness: negative feedbacks, high cooperativity of transcriptional regulation (thus switch-like behavior, changing [TF] may have little effect on target expression as long as it is above some threshold values).

3. Origin of robustness [Wagner, 2005, Chapter 13, 16, 17]

Definition of robustness:

- Robustness: for a biological system with many possible states, mutations (leading to change of states) do not change significantly the function/phenotype of the system. Examples: (1) protein: function is robust to change in protein sequence; (2) CRS: expression pattern is robust to change in sequence; (3) metabolic pathway: flux is robust to change in enzyme activity; (4) metabolic network: input-output function is robust to change of individual reactions.
- Qualificiation of robustness: the notion that lack of change in the function/phenotype of a system needs to be qualified. The mutations may still have some other functional consequence (see below), so robustness only refers to certain well-defined aspect of the system that is robust to mutations under a specific environment and genetic background.
- Multi-level nature of robustness: even changes of phenotypes in one level may not lead to change of fitness if the upper-level system is robust to changes. Example: (1) protein: change sequence → change enzyme activity, but that does not change the metabolic flux, as the pathway is robust to individual enzymes; (2) CRS: change sequence → change expression pattern, but that does not change the development process, as the GRN controlling the relevant development process is robust to change of expression of individual genes.
- Interpreting evidence of robustness: in some cases, it is not clear whether there is truely robustness or at which level robustness operates. Example: (1) CRS: different orthologous CRSs drive the same expression in different species. Could be due to compensatory change of trans- environment; (2) metabolic network: removing enzymes do not affect growth. Could be due to the inability of detecting fitness changes at different conditions.

Robustness as adaptation to mutations:

- Neutral space: for a given phenotype/function, the collection of all solutions in the state space.
  - Example: all proteins that perform the same catalytic function.
  - Heterogenity of neutral space: not all solutions in the neutral space are identical, they may still differ in some other aspects. Example: (1) proteins: may differ in structure; (2) regulatory sequences: may differ in the composition of TFBSs; (3) regulatory networks: may differ in the design of networks (e.g. the number of feedback loops). As a consequence, some regions (defined by some other aspects) contain many more states than other regions (frequent vs rare regions). We can thus define robustness at the level of regions.

  - Robustness at the level of states: some states have more neighbors in the neutral space, and we call these the robust states.

- Intuition of evolution of robustness: it is much more likely to find a solution in the frequent regions from a blind search by evolution; and similarly, as evolution proceeds, it is more likely to move into regions with more states/neighbors on average. A simple example: two structures with different robustness in the same population. At each generation, the low-robustness structure tends to produce more deleterious mutations than the high-robustness structure, thus over time, the frequency of low-robustness structure will be reduced in the population.
- Analysis of evolution of robustness:
  - Large population size: $N\mu >> 1$, the population is polymorphic most of the time. Consider a neutral network G of multiple states, then the population exists at a mutation-selection balance over all possible states in G. The frequency of each state in the population can be computed for the steady-state: the highly connected states (robust states) have higher frequencies.
  - Small population size: $N\mu << 1$, the population is monomorphic most of the time. The behavior of the system on a neutral network can be understood as the process of random walk in a space (Markov chain): the equlibrium probability of a region is proportional to the number of states in that region, if each state is equally likely (from detailed balance of Markov chain).
  - Mutation rate: not the same as DNA mutation rate, e.g. for a DNA sequence of length L, it would be $L\mu$, where μ is the per bp mutation rate. Therefore, for large genetic networks, the mutation rate is significantly higher, and the condition $N\mu > 1$ is easier to meet.
  - In general, if evolution is not strictly neutral (under adapative selection or allow nearly neutral changes), then the probability distribution for small populations, or the steady-state frequencies for large populations, depend on both the rate of change among states (dependent on fitnessness), and the number of states in each region.

Robustness as adaptation to non-genetic changes/noises: by-product of selection for non-genetic robustness.

- Non-genetic changes/noises: environmental variations. Gene expression noises: affect individual gene expression and also phenotypic level, e.g. wing pigmentation pattern in fruit fly.
- Robustness to nongenetic changes entails mutational robustness: examples (1) protein/RNA molecules: robust to thermal noises; (2) Hsp90: protect proteins for heat shock, also make protein folding robust to mutations; (3) segment polarity network: robust to gene expression noises and other variations, e.g. embryo body size.
- Natural selection can modify robustness to nongenetic changes: this non-genetic robustness is highly dependent on genetics. Examples: recA mutant in E. coli has much higher expression noises; in fruit fly, Hb has a much lower expression noises than other genes such as Bcd, and this is controlled by other genes (staufen may be one). Also note that non-genetic mutation is an important driver of selection: the variance from non-genetic factors is usually much larger than the variance from mutations: $V_m << V_e$.

4. Mechanism of robustness

   Reference: [Wagner, 2005, Chapter 15]

   Gene duplication contributes, but is not a main source of robustness:

   - Cases of gene duplication and robustness to mutations: these genes have duplicates and no phenotype if deleted: (1) HMG2: enzyme in sterol biosynthesis (important for many other pathways from electron transport to DNA repair), HMG1 is a duplicate; (2) CLN1: regulator of Cdc28, two duplicates CLN2 and CLN3; (3) TPK: signaling genes, 3 duplicates; (4) knirps and knirps-related: head development.

- Gene duplication is not a main explanation of mutational robustness of yeast genes (deletion does not have phenotype): no/weak correlation between number of copies and phenotypic effects, between dispensability and protein evolution rate, etc.
- Rapid functional divergence of duplicated genes: diverge fast as measured by protein evolution rates; by regulatory control (in ChIP-chip data); by gene expression.

Distributed robustness: offers better explanation of robustness. The fundamental source of robustness comes from: the effect of change may be diluted or compensated by the other parts of the system. Some general mechanisms:

- Distributed functions: distribute the function of a system into multiple components, s.t. each plays a small role. Ex. metabolic pathways: multiple enzymes; protein folding: multiple AA interactions contribute to the protein stability.
- Feedbacks and cross-talks: this helps compensating for a change/variation through changes in other parts of the system. Ex. two pathways controling the same genes, and when one pathway is shut down, it sends signal to the other pathway, which may take over the role.
- Steady states: some features may allow the system to have strong steady states, relatively insensitive to changes. Ex. positive feedbacks may help "lock-in" a state, contributing to the steady states.

Programming and checkpoints::

- Principle: a system may enter a state (through "programming") that is "simpler" and thus insensitive to many changes/perturbations.
- The best example is cell differentiation, through committing to a state (fate), part of the cell is essentially shut down and will no longer be responsive to many external signals. This is implemented in epigenetic control (as if DNA sequences are removed after cell fate committment).
- Similar example may be cell cycle control, that has clear check points.
- Normal cells are protected against cancer, being a good example of robustness. Part of the protection is encoded in cellular differentiation (proliferation limited, unable to migrate to other tissues, etc.).

Mechanisms of robustness in various examples:

- Protein structure: many sequences often fold to the same structure, from the interactions of many non-redundant blocks. Example: change one AA, could fold slightly different to accomodate the different AA.
- Metabolic pathways: the flux is determined by multiple enzymes in a pathway.
- Metabolic networks: reroute the flux when one pathway is blocked.
- GRN in development: negative feedbacks, signaling cross-talk, etc.

5. Robustness and evolvability [Wagner, 2005, Chapter 14]

The qualified view of neutral changes: a neutral change may not be strictly neutral, as it still changes some other aspect.

- Multi-functionality: well-known for proteins. Examples: (1) phosphoglucose isomerase (glycolysis): also a cytokine of immune system; (2) aminoacyl-tRNA synthease: also regulate transcription and translation; (3) segment polarity regulators: may also regulate development of other regions at different stages.

- Different functions at different conditions or genetic background: Examples: (1) synonymous substitutions: can be phenotypically different depending on codon usage; (2) metabolic enzymes: mutations may only affect fitness under some conditions.

Neutrality and innovation:

- Evolvability: the system can acquire novel functions through genetic changes.
- Neutral changes (in some aspects) can lead to innovations in another aspect (adapation), or at a later point when environment changes (ex-adaptation). Examples: (1) metabolic enzymes acqure new functions: lactate dehydrogenase as crystallin in eyes; (2) segment polarity gene acquire new function in eyespot formation.

6. Limitation of robustness [Wagner, 2005, Chapter 18]

   Problem: some biological systems are NOT robust. What limits the robustness of these systems? Example: some enhancers are sensitive to minor changes of sequences.

   Limitation of robustness:

   - Variations: increasing mutation rates can create more variations and thus reduce robustness.
     - Examples: bacterial mutation rate is incresaed during stress. One mechanism is to suppress DNA repair. Another example is Hsp90, which is a capacity of genetic variation.
     - However, increasing genetic variation may be a by-product of something else: e.g. DNA repair is costly and thus better to avoid in stress; Hsp90 is mainly used for helping folding of proteins. It is hard to evolve mechanisms that increase variations under individual-based selection.
   - Trade-offs with other features/aspects of system: this is probably the main mechanism that limits robustness.
     - Overlapping genes: common in virus. This feature reduces the robustness to mutations, but increases the DNA replication speed, thus may provide advantages.
     - Protein/RNA structure: a trade-off between thermo stability and function (which may requires less robust structure).

1.3 Evolution of Biological Systems: Theoretical Studies and General Principles

What influence the evolution of a complex biological system?

- Structure of the fitness landscape or the neutral space. This determined by the system features including: stability of components, redundency, and distribution of control.
- Population size and mutation rate: basic population genetic parameters.
- Environmental variation.
- History of evolution: chance and necessarity.

Principles of evolution of complex biological systems:

- Error threshold: if mutation rate is very high, the population cannot sustain its optimal phenotype - “survival of flattest”.
- Genotype-phenotype map is usually highly uneven: some phenotypes correspond to many genotypes.
- Robustness of a system may come from: redundancy, distribution of control, etc.

- Adaptation to non-genetic changes may increase robustness.
- Robustness of a system can facilitate adaptation by: (1) random exploration of a larger neutral space; (2) ex-adaptation.
- Evolution is often determined by trade-offs: e.g. protein function (benefit) and expression (cost); overlapping genes replicate fast (benefit) but is more constrained (e.g. in gene regulation, cost).
- Complexity of a system may come from simple processes, or non-selective forces [Lynch]. Ex. in small populations, selection is not effective, thus allow genomes to grow and get more complex.

Does selection mold molecular networks [Wagner, Sci. STKE, 2003]

- Problem: why does a biological network has certain structure?
- Systems:
  - Well-characterized systems: MAPK signaling pathway, lyosegny-lysis switch of bacterialphage, segmentation genes in fly, flower development genes in plants.
  - Large-scale networks: topological features. Detailed modeling is generally infeasible.
- Methods:
  - The behavior of the system that may confer some advantages: e.g. ultrasensitivity of MAPK pathway (switch-like behavior).
  - Statistical patterns of network topology: e.g. the overrepresentation of certain motifs.
  - Other evolutionary signatures: e.g. the constraint of hub vs other proteins.
- Remark: when making claims about a design (under selection), it is important to consider alternative hypothesis: (i) alternative designs that could also achieved certain properties (e.g. sensitivity of MAPK pathway can be achieved via multiple phosphorylation sites); (ii) other processes that may explain the feature without invoking selection: e.g. protein duplication that leads to scale-free network topology.

The nature of neutrality [Wagner, FEBS Letters, 2005]

- Hypothesis:
  - A mutation that does not change one aspect of a biological system's function may become non-neutral in a different context, thus neutral muation should always be defined wrt. some aspect in a specific environment and genetic background. In other words, a neutral mutation could be a source of innovation when the context changes.
  - A robustness system harbors more neutral changes, thus is capable of exploring more configurations, and evolve more innovations.
- Examples/cases of neutral mutations in one aspect/environment/time becomes non-neutral in a different one:
  - Environmental dependence: for an enzyme of some carbon source such as gluconate, a mutation may be neutral if the environment is dominated by other carbon sources, but not if gluconate is the sole carbon source.
  - Genetic dependence: many common genetic diseases, the same mutation may cause severe disease in one individual, but no effect in another.
  - Multi-functionality of genes: a gene may serve multiple functions, e.g. phosphoglucose isomerase is also neuroleukin, a cytokine causing immune cell maturation. Thus a mutation that does not affect one aspect of function may change another aspect.

  - Multi-stage/tissue expression of genes: a gene may be expressed/used in different stages/tissues, thus a mutation (cis-) that does not change its expression in one stage/tissue may change another.
  - RNA structure: two positions x and y, $C \to G$ at x is neutral if $y = G$, but not if $y = A$ (Figure 1).

Balancing robustness and evolvability [Lenski & Ofria, PLoS Biol, 2006]

- Mechanisms of robustness:
  - Stability of individual components: e.g. stability of protein structure/function, cooperativity of interactions between proteins and ligands/DNA (switch-like behavior), etc.
  - Redundency of component parts.
  - Distributed robustness: e.g. negative feedbacks.
- Origin of robustness:
  - Mutational robustness can arise from: high mutation rate, or sex. "Survival of the flattest": at high mutation rate, most offsprings carry mutations and selection favors populations that find lower fitness peaks surrounded by less precipitous mutational chasms.
  - By-product of selection for robustness in the face of variable environments.
- Robustness and evolvability: robustness may increase evolvability through:
  - In the case of redundency: promoter adapatation by allowing duplicated genes to evolve new functions.
  - Neutral network provides evolutionary paths to new adaptations by random drift, in effect allowing populations to search for wider regions of genotypic space for rare beneficial mutations.

Robustness and evolvability [Wagner, PTRSB, 2008]

- Problem: does robustness of a system enhance its evolvability?
- Defintions:
  - Sequence robustness: the number of neutral neighbors of G.
  - Sequence evolvability: the number of different phenotypes in the neighborhood of G.
  - Structure (phenotype) robustness: the number of neutral neighbors of this phenotype, P.
  - Structure (phenotype) evolvability: the number of different phenotypes in the neighborhood of P.
- Results:
  - High sequence robustness means low evolvability: more neighbors are neutral, thus not many different phenotypes.
  - High phenotype robustness means high evolvability: larger neighborhood because of high robustness, thus possess more phenotypes.
  - More robust phenotypes can acess more variation in their evolution on a neutral network: in both cases where $N\mu >> 1$ (polymorphism in one generation) and $N\mu << 1$ (accumulative variation across many generations).

Evolution of spatial expression pattern [Johnson & Brookfield, Evolution & Development, 2003]:

- Problem: the factors that influence the adaptive evolution of novel expression patterns; and the properties of the solutions (networks) evolved for certain target patterns.

- Methods:
  - Gene regulatory networks: 11-cells, the trigger (initial morphogen gradient), the target and other intermediate TFs. For each cell:

    $$[S]_{t+1} = 0.1 \prod_{R=1}^{n} effect_{R,S}^{[R]_t} \tag{1.5}$$

    The expression level is always between 0 and 1. And the final TF concentrations are obtained after 100 time steps.
  - Fitness/success of a network: define the desired expression patterns of the target gene: posterior or extremes (in both ends), where each value is 0 or 1. The deviation D is defined as the sum of the absolute difference at each cell. And the success is: $1 - D/D_0$, where D_0 is the initial deviation.
  - Evolutionary simulation: fixed network topology, each mutation affect the regulatory effects randomly. Apply 1000 mutations and assess the success of the network. Repeat 100 times from the same starting point.
- Probability of success: generally low, e.g. only 5% for 1-intermediate network for the extremes pattern.
  - The complex networks are only slighly better than simpler networks.
  - If allowing slightly deleterious mutations, then complex networks have greater success in reaching the favored pattern while the simpler networks not. This could be viewed as capturing the effect of population size.
- Properties of the solutions:
  - Some similarities in evolved solutions: e.g. in 2- and 3-intermediate networks, most successful solutions use one intermediate gene as an antagonist of the target.
  - Partial redundency: for complex networks, the evolved solutions often have partial redundency, i.e. if removing one or more genes, the result pattern is not significantly affected.

Evolutionary tuning of gene expression [Dekel & Alon, Nature, 2005]:

- Hypothesis: the expression level of a protein is under evolutionary tuning/optimization.
- Model: consider the case of lactose utilization. Producing the enzymes incurs cost, and the utilization of lactose by the enzymes produces benefits. The cost and benefit can be expressed as the relative growth rates of cells.
  - Cost: a function of Z (level of lacZ) (more protein, more cost). Assume the form: $\eta(Z) = \eta_0 Z/(1 - Z/M)$, where M is the limit of Z.
  - Benefit: the increse in growth rate is due to the increase of lactose use. A function of Z and L_{in}, the intracellular lactose concentration (more lactose and more utilization, i.e. more enzyme, more benefits). Following lactose transport and the cataboism kinetics: $B(Z) = \delta[ZL_{in}]$, where δ is the growth advantage per lacZ molecule at saturating lactose concentration.

  The parameters are fit by experimentally measuring cost and benefit.
- Methods:
  - Experimental evolution: serial diluation assay, 1:100 dilutation everyday, i.e. $\log_2 100 = 6.6$ generations per day.
  - Measuring cost: induction of lacZ by IPTG, thus no actual utilization of lactose (no benefit) and measure growth. Vary Z by using different IPTG levels.

  - Measuring benefit: fix the cost by using saturating IPTG (thus full induction of lacZ), and vary lactose level and measure growth.

- Results:
  - Putting E. coli at different lactose level, and observe the change of lacZ: lacZ reaches the optimal (theoretically predicted) level within 300 - 500 generations. The dynamics can be explained by a single mutation (about 1 in 100 possible mutations can fix the lacZ expression)

- Remark:
  - The difficulty in this experiment is to seperate cost and benefit, which are coupled in normal conditions. The idea is to control one of them.
  - The cost-benefit analysis: the key element of such analysis is to have a uniform measure of cost and benefit, e.g. for cost-effectiveness analysis of treatments, use Quality Adjusted Life Year (QALY) to measure the overall effectiveness.

Evolution under varying goals [Kashtan & Alon, PNAS, 2007]:

- Problem: will time-varying goals speed up evolution?

- Methods:
  - Model systems: logic circuits (using basic building blocks, NAND gates, to implement complex Boolean functions), RNA, etc.
  - Simulation procedure: standard genetic algorithm. Fitness is defined as the fraction of all possible input values for which the network gives the desired output.
  - Schemes of evolution:
    - Modular varying goals (MVG): suppose g is the goal, then g' is a variation of g if g' shares some components with g, but different in some other components. Ex. $g = (xXORy)AND(wXORz)$, then a varying goal is $g' = (xXORy)OR(wXORz)$, where one gate AND is replaced by OR, but other subgoals are unchanged.
    - Random varying goals (RVG): g' is another randomly selected goal. RVG_v where the random goal is randomly selected each time; RVG_c where the random goal is fixed.
    - VG_0: varying between g and neutral evolution.

- Results:
  - MVG significantly speeds up evolution: 50 130 times faster for logic circuits; and 20 30 times for RNA. The speedup is greater for complex goals. RVG_v also speeds up for logic ciruits, but not for RNA. RVG_c and VG_0: no speedup.
  - Speed up is robust to simulation parametes (time of changing the goals, population size, etc.); and also held for hill-climbing algorithm instead of genetic algorithm.

- Discussion: the interpretation/intuition is (Fig. 5): under fixed goal, the population is often trapped in local maxima for long time. Under MVG, the population, when in a local maxima, will be pushed away with a different goal; meanwhile, since the new goal is similar to the original goal, the population will still stay in the neighborhood.

Phenotypic switching [Bistability in feedback circuits as a byproduct of evolution of evolvability. MSB, 2012]

- Modeling stochastic chemical systems: stochastic chemical kinetic equations (SCK). The naive Stochastic Simulation Algorithm simulates reaction events one at a time, not very scalable to large systems. The idea is to have higher-level abstraction: e.g. approximating fast reactions; mix of determistic and stochastic treatments, etc.

- Environmental fluctuation and stochastic expression: how env. fluctuation influences/selects the stochasticity of gene expression?
  - The model system: transcription of gene G with positive feedback. The system achieves bistabability at large N (Hill coefficient).
  - Environment: define fitness functions on high, and low levels of G. And the fitness requirement changes over time.

  The results suggest that stochastic expression is preferred at intermedidate rate of env. fluctuation, and high nonlinearity of the system.

Mutational robustness and evolvability [Draghi & Plotkin, Nature, 2010]

- Hypothesis: mutational robustness can facilitate adaptation. The intuition is: neutral mutations can lead to genotypes with very different phenotypic neighborhoods, thus will allow them to adapt to novel phenotypes.
- Methods: a population genetic model, N individuals, with P possible phenotypes. For any individual, the number of accessible phenotypes is K, and μ is the probability that a mutation creates a new genotype. With probability q, the mutation is neutral, thus q is a measure of robustness.
- Results:
  - Very large q: reduce the phenotype variation (i.e. most mutations will be neutral, thus hard to evolve differnet phenotypes, or speaking in other works, trapped in the current phenotype); very small q: most mutations will be deleterious, thus the population can only evolve within a very space, hard to reach different phenotypes. The best q is an intermediate value. Note: the evolvability is measured by (1) the time to evolve a new phenotype; (2) the phenotypic diversity of the population (assume $N\mu > 1$, thus the system is at mutation-selection balance).
  - Population size and mutation rate have important effects.

1.3.1 Evolution of Biological Molecules

Evolutionary dynamics of biological sequences in fitness lanscape [Schuster & Reidys, 1997]

- Problem: evolutionary dynamics of sequences. Similar to the study of dynamic systems, we could ask questions such as: whether steady states exist or oscillation; the time converging to the steady states; whether a population is capable of reaching the desired phenotypes; etc.
- Genotype-phenotype mappings (sequence-structure mappings of RNA sequences):
  - Concepts: shape space covering - the space of molecules to be searched to reach a given structure; neutral networks - all sequences that fold into the same structure.
  - Properties of the mapping: e.g. for RNA molecules, the number of sequences is much larger than the number of structures; individual structures differ greatly in their frequencies; sequences forming a structure are found be randomly distributed in sequence space; some neutral networks form a giant component in the sequence space.
- Evolutionary dynamics/optimization on landscapes:
  - Error thresholds: if the error rates exceed some threshold, the best structure is lost and the population drifts randomly in sequence space.
  - Alternating fixation of two structures: the overlap of the two structures in sequence space is crucial for transitions.

- Discussion: two assumptions by the model:
  - Constant environment/fitness: not applicable if, for example, two RNA molecules co-evolve.
  - Independent reproduction: fitness may be frequency-dependent.

Modeling evo-devo with RNA [Fontana, BioEssays, 2002]

- Problem: the properties of genotype-phenotype mapping and its implications on evolution of phenotypes, e.g. the rate, the directionaility of phenotypic changes, the constraints on whether a phenotype is evolvable, whether phenotypic changes are gradual or punctuated, etc.
- Main thesis: the statistical architecture of the sequence-to-structure map in RNA offers explanations for patterns of phenotypic evolution.
- Background on RNA phenotypes:
  - Secondary structure: loops and stacks. The major stabilizing free contribution contribution comes from stacking interaction between adjacent base pairs. Loops are destabilizing.
  - The secondary structure participates as a geometric, kinetic and thermodynamic scaffold in the formation of the three-dimensional structure, which involves bringing secondary structure elements into proximity by means of pseudoknots, non-standard base pairings and bivalent counter ions.
  - Phenotypic plasticity: a single RNA sequence wiggle between alternative low energy shapes, defined by Boltzmann distribution. The set of all shapes within a free energy interval (say, 5kT) from the ground state is called the plastic repertoire (a measure of plasticity).
  - Norm of reaction: the melting profile of a RNA sequence changes at different temperatures.
- RNA folding map and its evolutionary consequences:
  - One phenotypes, many genotypes: the number of structures is much smaller than the number of sequences, and the frequency of shapes is strongly biased.
  - Thermodynamic stability and combinatorial realizability: a frequent shape compromises between two opposing trends: while long stacks enhance the thermodynamic stability of a shape, they lower its combinatorial realizability by constraining the choice of nucleotides at paired positions.
  - Neutral networks and evolution of phenotypes:
    - Improving the chance of encounting new phenotpes: the population can drift on that network into far away regions, vastly improving its chances of encountering the neutral network associated with a different phenotype.
    - Evolution towards robustness: a selection/mutation balance on a neutral network automatically yields phenotypes that are relatively robust to mutations (non-robust genotypes are more easily lost because of mutations).
  - Shape space covering: for a random sequence, the average number of mutations to realize any frequent shape is much smaller than the radius of the sequence space (e.g. 15 mutations for a length 100 sequence).
  - Continunity and constraint of phenotypic evolution: define accessibility between two phenotypes as the adjacency of their corresponding neutral networks in sequence space (asymmetric relation). A path is continuous if the phenotype of the offspring is near the phenotype of the parent in the accessibility topology. For any two shapes, there is not always a continuous path between them. This is similar to “developmental constraints”.

The ascent of the abundant in RNA evolution [Cowperthwaite & Meyers, PLCB, 2008]

- Problem: what phenotypes are evolvable? Or will evolution always succeed in reaching the optimal phenotypes?

- Methods:
  - Fitness of RNA molecules: suppose σ is the structure of a molecule m, and t is the target structure, then the fitness of the molecule is given by:

$$W(m) = \frac{1}{\alpha + [d(\sigma, t)/L]^{\beta}} \tag{1.6}$$

    where $\alpha = 0.01$ and $\beta = 1$ are scaling constants, $d(\sigma, t)$ is the Hamming distance, and $L = 12$ is the sequence length.
  - Simulating evolution: a population of $N = 1000$, mutation rate $U = 0.0003$. Check if the target phenotype was evolved after $\tau = 10^6$ generations: it is considered successful if the target phenotype occurs in the population, regardless of its frequency in the population.
  - Defintions: abudance - the number of genotypes that produces a phenotype.
- Results:
  - Abundance of target phenotype and success of evolution: a significant positive correlation between the abundance of the target phenotype and the likelihood that a population successfully evolved to the target.
  - Abundance of founding phenotype and success of evolution: no significant correlation. When it fails, the population was often trapped in phenotypes of greater abundance than both the target phenptype and the average of random phenotypes.
  - Naturally occuring RNA molecules: biased toward abundant phenotypes.
- Conclusion: abundant phenotpyes may be easy to find, but difficult to escape. Thus, the evolution of phenotypes may be biased toward abundant phenotypes, even if those phenotypes are not optimal.

1.3.2 Network Evolution

Evolution of molecular pathways and networks [Cork & Purugganan, BioEssays, 2004]:

- Problem: how the roles of genes in the biological pathways and networks constrain/influence their evolution?
- Pathway-level evolution:
  - Genes early in pathways may be under stronger selection force than in downstream ones. Ex. plant anthocyanin biosynthesis pathway.
  - Genes early in pathways may also be targets of positive selection. Ex. plant floral developmental pathway. Possible explanation: the regulatory genes will change when and where the downstream gene batteries to be expressed.
  - Genes in the branch point of metabolic pathways can be targets of positive selection. Ex. glycolysis pathway in D. melanogaster and starch biosynthesis pathway in plant. The change of these enzymes may change the partition of fluxes.
- Network-level evolution:
  - Network growth (PPI): preferrential attachment vs addition of modules.
  - Network (PPI) hubs tend to evolve more slowly; the interacting genes tend to co-evolve, and have similar evolutionary rates.

Evolution of complexity in signaling networks [Soyer & Bonhoeffer, PNAS, 2006]:

- Problem: biological systems often show a level of complexity that is above the minimum required (e.g. from results of synthetic biology), why?
- Idea: many neutral changes tend to increase the complexity of systems (robust systems generally allow more neutral mutations).
- Methods:
  - Pathway model: a network of proteins, where each protein has active and inactive states. The proportion of the two states is determined by the effect of interacting proteins (a matrix of protein-protein interactions): the rate of active to inactive (or inactive to active) state is a linear function of the concentration of the active interacting proteins.
  - Evolutionary simulation:
    - Selection criteria: one protein in the network responds to the signal (receptor) and another one is the effector. So the behavior of the network is defined by the change of effector protein in response to the change of receptor. A certain type of behavior is desired in each case.
    - Fitness function: if the selection criterion is not met, fitness is 0; if it is met, fitness is $1 - nc$, where n is the number of proteins, and c is the cost of adding one protein (a small number).
    - Mutation events: mutations adding a protein (protein duplication) or an interaction (new interacting site); mutations deleting a protein or an interaction.
    - Simulation: start with a population of 1,000 random pathways consisting of 3 proteins. A new generation is sampled from the population in the last generation with replacement: the probability of sampling an individual is proportional to its fitness. And the population is subject to mutation events.
- Results:
  - Starting from minimum pathway (3 proteins): the average size of the population increases, and reach an "equlibrium".
  - Mechanism: when the pathway size is small, the deletion events are more likely to be lethal than insertion events. Until the pathway is big enough, the two types events balance each other.
- Conclusion: the complexity of a biological system may not be always caused by the need of increasing fitness.

Network designability [Nochomovitz & Li, PNAS, 2006]:

- Problem: are some phenotypes more "designable" than others, where designable means more ways of getting the same phenotype from different networks?
- Concepts:
  - Designability: the number of networks with the same dynamic phenotype.
  - Robustness: the basin of attraction of a phenotype (how many transient states will converge to the stable phenotype.
  - Mutational buffering: phenotypic robustness to internal genetic variation amongst different strains.
- Methods:
  - Phenotype characterization: typically limit cycles consisting of multiple states (a network may exist in multiple states, e.g. a 4-node network may exist in 16 states). The length of the limit cycle is the key parameter.

  - Network and phenotype mapping: Boolean network, where each node receives input from other nodes (activation or repression or no effect), the state of a node depends on the numbers of activations and repressions received in the last time point.

- Results:
  - Certain dynamic phenotypes are more designable: enumerating all 3-node and 4-node networks and map phenotypes. The distribution of designability suggests the overrepresentation of highly designable phenotypes.
  - Designability is indepdent of robustness.
  - Mutational buffering: observe that despite that different 3-node networks have different phenotypes, adding a 4-th node creates the same phenotype, thus mutational buffering (at the level of 4-node network).

Innovation and robustness in complex regulatory gene networks [Ciliberti & Wagner, PNAS, 2007]:

- Hypothesis: the degeneracy of network-phenotype space allows evolution of novel phenotypes.
- Methods: genotype - Boolean networks, phenotype - expression patterns of the network (at equilibrium).
- Results:
  - Genotype-phenotype space: highly degenerate, the neutral space of one phenotype can be vast (in genotypic distance).
  - Two networks with the same phenotype can evolve very different phenotypes, i.e. the phenotypes of their neighbors (or "phenotypic neighborhood") may be very different.

Evolution of metazoan segmentation [Francois & Siggia, MSB, 2007]:

- Problem: how complex traits, segmentation in this case, are evolved?
- Idea: evolution rewards increasingly more complex patterns: more segments will have higher fitness. Thus a complex trait can be evolved gradually.
- Methods:
  - Transcriptional regulation: activators - Hill type activation; repressors - reduce the activation (by activators) by a fraction, which is a Hill functino of repressor concentration.
  - Expression pattern of TRN: the morphogen (G) gradient (monotonic); the repressor (R) and effector (E) expression pattern is determined by their regulators in TRN.
  - Fitness of TRN: only determined by the effector expression: number of segments.
  - Mutations: add or delete genes; add or delete links.
  - Simulation procedure: start with 100 random networks. At each generation, the top 50 networks (by fitness) are chosen for reproduction and mutation.
- Results:
  - In general, successive steps of adding repressors s.t. the patterns of effectors are getting more complex (Figure 3). Ex. the process of evolving the pattern of E with a single stripe (Kr): repressor $R1$ of E, then another repressor $R2$, which represses $R1$ and is activated by G.
- Remark: the evolution still requires multiple events: e.g. a new repressor $R1$, which supppresses E, and is regulated by G.

Selection of alternative mode of gene regulation [Gerland & Hwa, 2009]:

- Problem: genes whose expressions are induced by ligands can be regulated through different mechanisms. Double positive control: inducer → transcriptional activator → target expression; or double negative control: inducer deactivates transcriptional repressor, which without inducers will suppress expression. Which mode of regulation will be chosen?
- Observations: the mode of regulation is determined by how often the gene is used (the demand of gene). High demand: the gene is often needed, i.e., the expression level (ON) is under selection most of the time. Two competing possibilities:
  - Use-it-or-lose-it: the gene should be used most of the time, thus it (the mechanism that leads to gene expression) will be maintained when the gene is not needed (shorted time period).
  - Wear-and-tear: the usage of the gene should be kept minimum, so that the mutation will have minimum detrimental effects (minimize the time when the mutation will be harmful).

  Which one will dominate?
- Model:
  - The description of the system:
    - A population has two alleles: wide-type or binders, and mutants or non-binders. The non-binders cannot bind to DNA. For double positive (activator) system, the non-binder is always OFF; for double negative (repressor) system, the non-binder is always ON.
    - The inducer level alternates between high and low states: at high state, demand high gene expression; at low sate, demand low gene expression. Thus, for activator system, non-binder has reduced fitness at high state; for repressor system, non-binder has reduced fitness at low state.
    - Different modes of regulation: different fitness (assume there are multiple populations with different modes that compete with each other). The mode with higher fitness will be chosen.
  - Evolutionary dynamics when N is large: the frequency of non-binders $x(t)$ is determined by mutation rates between binders and non-binders, and selection on non-binders $s(t)$ (fitness reduction). The comparison between two modes is based on average fitness reduction:

$$\langle \gamma \rangle = -\frac{1}{T}\int_0^T s(t)x(t)dt \tag{1.7}$$

  - Evolutionary dynamics in finite population: need to consider the effect of random sampling. Wright-fisher model with population size N, and periodic selection.
- Analysis:
  - Behavior of system: for both systems, the fequency of non-binders, $x(t)$, changes according to:
    - In the selection phase (T_s): non-binders are eliminated, and under the assumption $sT_s >> 1$, the elimination is fast.
    - In the neutral phase (T_n): non-binders accumulate because of mutation.
  - Infinite population: assume high state is infequent (0.05), two cases
    - Double positive system: selection phase is short, and no fitness reduction in the long neutral phase. Thus overall fitness reduction is small.
    - Double negative system: selection phase is long, any mutation during selection phase will leads to fitness reduction. Thus overall fitness reduction is large.

    The result is: double positive system is preferred - “wear-and-tare”.

- Finite population: two cases
    - Double positive system: because of random drift, there is a chance that all binders are lost during the long neutral phase. Thus average fitness reduction is high.
    - Double negative system: opposite case, lower fitness reduction.

  The result is: double negative system is preferred - "use-it-or-lose-it".
- What determines which mode is chosen? Popoulation size is a critical parameter. To find it, approximate by a two-state system: the population switches between $x \approx 0$ and $x \approx 1$.

- Remark:
  - Design choice and evolution: the design choice a biological system makes may be determined by its evolutionary consequence. Two design choices may be equally good, if only look at the extant system and functionaility, but have different evolvabilities. To compare evolvabilities, one can assume that both choices are made (in different groups), and see which one has the higher fitness.
  - Evolutionary robustness: an important aspect of evovability is whether a design will persist (robustness in the evolutionary context) over time, in the face of other mutations, change of selection, etc.

Evolution of evolvability in GRN [Crombach & Hogeweg, PLCB, 2008]:

- Problem: will a population of biological systems become more evolvability (more likely to have beneficial mutations) over time? Or is evolvability iself under selection?

- Idea: giving a changing environment (e.g. the alternation between two evolutionary targets), some networks can more easily switch between different targets, i.e. more evolvable. These networks will ultimately dominate in the population, thus the evolution of evolvability.

- Methods:
  - Network model: each individual is represented by a genome: a linear array of genes and binding sites. The neighboring sites of genes determine the GRN. And the dynamic behavior of GRN is modeled by Boolean networks (supposedly, the steady state expression patterns as the phenotypes of GRNs). Note that each gene has an expression threshold and the expression pattern is defined by ON and OFF states of genes.
  - Mutational events: (i) for genes: duplication, deletion and threshold change; (ii) for binding sites: duplication, deletion, innovation, weight change (activator to repressor or vice versa) and preference change.
  - Fitness and reproduction: the fitness of an individual is a monotic function of the Hamming distance between the expression pattern of the individual and the evolutionary target. The population exists in a spatial grid and an individual is able to generate offsprings in the neighboring empty grid cells in each generation, whose probability depends on the fitness.
  - Evolutionary targets: two randomly chosen expression patterns, alternating over time.

- Results:
  - Evolution of evolvability: at the initial stage of evolution, the adapation is slow and the network is unable to completely adapt. The adapation time, defined as the time of responding to a new target s.t. the Hamming distance to the new target reduces to 0, is much shorter later in evolution.
  - Mechanism of quick adaptation (evolutionary sensor, ES): duplications of a few genes are important for quick adaptation. It was found that changing copy number of gene 3 and 6 (both are hub genes, well-connected), could quickly change the phenotype of GRN s.t. the new target can be approached quickly.
  - Mutational robustness: maintained throughout evolution, i.e. most mutations are neutral.

Chapter 2

Comparative Genomics

Problems of genetics and comparative genomics: the central theme is to explain intra- and inter- species variations of genetic features, traits, and associate them with environmental or internal changes:

- Genetic vs non-genetic influence of traits: the contribution of each, and in particular, how they interact to affect traits.

- Patterns of evolutionary changes: how often are the different types of changes? Correlation with divergence time?

- Natural selection on genotypes and traits: Are the evolutionary changes (or conservation) mainly driven by selection or random drift/mutation (neutral)? Role of negative and positive selection?

- Ultimate causes affecting the evolution of a biological system: Constraits that limit the change of traits: eg. if two proteins interact, their interacting residues may co-evolve. Or the adaptative force that drives the changes: e.g. changing environment (say cold) drives the changes at genome.

- Immediate causes (molecular mechanisms) of evolution of traits: for expression changes, cis- vs. trans-? For proteins, changes at critical sites?

- What does evolution tell us about how biological systems function? Ex. what types of changes are tolerated. We answer this question through the analysis of pattern of selection (changes may not be due to selection, and they are not informative for the true constraints of the system).

- Relation of molecular-level evolution to phenotypic evolution: does one level of changes always lead to changes at the next level? This also provide ultimate explanation as to the driving force of negative/positive selection.

Methods of genetics and comparative genomics:

- Genetic mapping: explore phenotypic variations through (natural) genetic perturbations.

- Conservation: if some trait or genetic feature is more conserved than expected by chance, then it is likely under negative selection. Note that to translate conservation to negative selection requires some caution, as conservation could be due to shared ancestry.

- Divergence: the higher divergence than expected by chance may suggest positive selection.

- Co-evolution of genetic features/traits: simply speaking, if two traits are correlated across evolutionary time, then it may reveal some relationship between the two, e.g. interacting proteins, or two traits that are determined by the shared genetic basis, etc.

- Association of genetic features, traits and other factors (environmental changes): e.g. the expression trait correlated with phenotypic change may suggest the gene whose expression is important for that phenotype.
- Infer the underlying knowlege of biological systems from evolutionary patterns: ex. infer whether an enhancer sequence tolerates indels by testing if indels are allowed in evolution. Or infer sequences whose mutations may be deleterious.

Examples of applying the conservation analysis:

- Conservation of sequences to identify functional elements.
- Conservation of gene relationship (e.g. spatial clustering or co-expression) indicates functional modules
- Amino acid evolutionary pattern of a protein reveals selective constraints at different regions of the protein.
- How constrained the expression pattern of a gene is reveals the functional role of the gene.

Lessons of comparative genome analysis:

- The genetic features and trait could be: sequences, expression data, and higher-order features such as TFBS clustering/co-occurrence or gene coexpression.
- Different levels of analysis: the gap between sequence and phenotype can be bridged through intermediate layers including candidate genes of the phenotype; or molecular phenotypes (e.g. brain imaging data in the case of cognitive behavior).
- Detecting selection: by choosing appropriate cotrols, one may be able to avoid the confounding factors such as mutation rate, effective population size, etc.
- Not all changes are adaptive: need to consider the impact of mutation, random genetic drift, recombination, population structure/migration to infer selection.

2.1 Basic Comparative Genomic Methods

2.1.1 Models of Molecular Evolution

Reference: [Yang, Computational Molecular Evolution, Chapter 2]

Empirical substitution models of amino acids:

- Empirical models: the general time-reversisble model:

  $$\pi_i q_{ij} = \pi_j q_{ji} \tag{2.1}$$

  It has $19 \times 20/2 - 1 = 189$ relative rate parameters and 19 free AA frequency parameters (total 208 parameters). Write $Q = S\Pi$, where S is symmetric, called AA exchangability matrix. The parameters can be estimated from a large number of protein sequences: DAYHOFF, JTT matrices. The matrix in mitochondrial is different because of different codon usage.
- Exchangabilities: depend on both the codon distance and on the chemical similarity of AAs. The former is very important, e.g. R and K codon differ only 1 nucleotide in genomic proteins and are chemically similar, thus high exchangability; but differ in two or three positions in mito. codon table, thus low exchangability.

Codon substitution models:

- Rate matrix: the subsitution rate depends on the nucleotide mutation rate, which is captured by HKY85 model; and the constraint on the AAs, modeled by the parameter ω.

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at two or three codon positions} \\ \pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \omega\pi_j & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition} \end{cases} \tag{2.2}$$

  ω is the commonly used d_N/d_S ratio: where d_N and d_S are the number of nonsynonymous/synonymous substitutions per site.

- Equilibrium codon frequencies: usually not empirical counts, which would need a lot of data. Use the product of the frequencies of the 3 nt that make up the codon ($F_1 \times 4$), or the product of position-specific nt frequencies ($F_3 \times 4$).

- Limitation of the codon substitution models: it is possible to model the physicochemical properties of codons, but in practice, this does not seem to improve much on the tasks such as positive selection. We may be limited in our understanding of how AA properties affect their evolution.

Random sites model: rate variation across sites:

- Random rates: to model rate variation, a common approach is to treat rates as RVs sampled from a prior distribution. The gamma distribution is commonly used with two parameters α and β. α is the shape parameter, typically, $\alpha > 1$: bell-shaped; $\alpha < 1$: highly skewed L-shape, and small α suggests higher rate variability.

- Continuity of rates: this may not be true. Ex. in MHC sequence, the positively selected sites are clustered in 3D structure, but scattered in primary sequence [Yang, Figure 8.5].

2.1.2 Phylogeny Reconstruction

PHYLIP:

- Maximum likelihood (ML) and maximum parsimony methods: protein and DNA sequences. For protein sequences, Dayhoff, JTT and other empirical models are used (but not codon models). For DNA sequences, JC, Kimura, F81, HKY85, etc. are used (again, not distinguish synonymous and nonsynonymous changes). Rate variation across sites can be included (s.t. synonymous rates may be captured to some extent).

- Distance methods: construct the tree according to the (pairwise) distance matrix. Methods include FITCH, Neighboring-joining, etc.

Phylogeny reconstruction in practice: in practice, both protein or DNA sequences are used (though DNA sequences seem to be common), the meaning of branch length may vary: e.g. number of subsitutions per site, or number of synonymous substitutions per codon.

- Yeast [Rokas & Carroll, Nature, 2003]: DNA sequences, codon alignment with ClustalW, and tree reconstruction by PAUP with ML or MP on 106 orthologous genes.

- Drosophila [Heger, Ponting, GR, 2007]: DNA sequences, tree reconstruction by FITCH program of PHYLIP - distance measured by the median d_S of about 6,000 orthologous genes.

- Yeast [Tuch & Johnson, PLoS Bio, 2008]: protein sequences, tree reconstruction by TREE-PUZZLE with VT model (empirical AA model).

SATe: simultaneous alignment and tree estimation [Liu & Warnow, Science, 2009]

- Goal: simultaneous alignment and reconstruction of the phylogenic tree.
- Methods:
  - Main procedure: starting with a tree T and alignment A, repeatedly: (i) construct alignment from the new tree; (ii) learn a new tree from the new alignment.
  - Tree construction: from the current alignment, use ML method RAxML.
  - Alignment: CT-i decomposition (divide-and-conquer method), create subproblems (subsets of sequences), then align them with an alignment tool; and progressively merge the subalignments.

2.1.3 Homology Mapping and Alignment

Goal:

- Partition of all genes in multiple species being analyzed into multiple families (orthologous groups).
- Infer the history of gene duplication and loss within each family.

Patterns of orthology: the analysis methods rely on these patterns to infer partition and history:

- Basic strategy: bidirectional best hits (BBH) are generally considered orthologs.
- Gene distance (similarity) should be consistency with the species tree: i.e. two genes in two species closely related in the species tree should also have similar sequences.
- History of gene duplication and loss should also be consistent with the species tree: this can used for distinguish in-paralogs and out-paralogs (in-paralogs should be closer).
- The number of gene level events (duplication and loss) should be small.
- Remark: the best strategy of infering both the partition and history thus should be simultaneous inference of both [SYNERGY, Bioinfo07], as one could help inference of the other.

Reference: [Gabaldon, GB, 2008]

HMM method to recognizing homologous proteins of a protein family [Eddy, HMMER user manual; Barrett & Karplus, CABIOS, 1997]

- Procedure:
  1. construct multiple alignment from given proteins of the same family
  2. train profile HMM (all parameters) from the aligned protein sequences
  3. database search: scoring each sequence in the database
- Scoring methods [DEKM, Biological sequence analysis, Figure 5.5, 5.6]
  - Observation: the raw log-likelihood score of the sequence depends on the length. Thus need to be normalized.
  - Z-score: for each length, collect all data points in the database and fit the mean and standard deviation. The score of a sequence is the z-score for that length.
  - log-odds ratio: protein family model vs the null model. This is length-normalized.
- Chooing null model for the scoring: the options are:
  - General protein-family independent null model: the aa distribution over all protein sequences in the DB

  - Protein-family specific null model: use the aa distribution of the specific protein family being tested. Average emission probabilities of aa's weighted by the transition probabilites of the states
  - Sequence-specific null model: train a null model for the sequence being tested

- Results:
  - The general model: the compositional bias is not captured. Thus if a sequence has similar composition as the (+) model, but different from H0, then it may reject H0.
  - The sequence-specific null model is probably too pessimistic: it requires (+) model to have a very good fit.
  - In HMMer, two null hypothesis are used: both the general one and the family specific one. They are combined by a Bayesian kind of posterior probability.

- Statistical significance of the database search:
  - Log-odds score (bit score): should be compared with log(N) where N is the database size. Generally, if it is larger than log(N), then it is significant.
  - E-value: the expected number of false positives in the database.
  - Calibrate the distribution of log-odds ratio: sample random sequences from H0, compute the score for each sequence, and do ML fitting of extreme-value distribution for the histogram of scores of sampled sequences (doing the local alignment, thus the best score asympotically follows EVD).
  - Correction for multiple testing: multiply by the database size
  - Empirical score cutoff: (i) trusted cutoff: the lowest score of the (+) sequences; (ii) noise cutoff: the highest score of the (-) sequences.

Handel [Holmes & Bruno, Bioinformatics, 2001]

- Aim: construct alignment from given multiple sequences.
- Idea: use a HMM to generate multiple alignment where one state of a HMM generates one possible alignment column.
- Methods:
  - Model: TKF91 indel model, represented as a pair HMM at each branch (Fig. 8), and Multiple HMM in a tree (factorize the probabilites). See Fig. 9 and 10 for Multiple HMMs corresponding to two simple trees (no automatic construction of Multiple HMM is described).
  - Inference: Sample multiple alignments by MCMC, in particular, to deal with internal nodes, use branch sampling and node sampling.

Transducer composition/PhyloComposer [Holmes, Bioinformatics, 2003; Holmes, Bioinformatics, 2007]

- Aim: automate the construction of MHMM (called evolutionary HMM, or EHMM) from the branch HMM structure and a guide tree.
- Methods:
  - Branch HMM (BHMM): two sequence transducer that takes an input sequence (parent node) and generates probabilistically an output sequence (child node). The mapping from input symbol to the output is encoding by the states of BHMM. A BHMM will have states:
    - ∗ Match states - M_{wx} which will emit symbol x with input symbol w;
    - ∗ Insertion states - I_x which will emit symbol x without input;
    - ∗ Deletetion states - D_w which will emit nothing for input w;

    - ∗ Wait state - W which will be needed for MHMM (indels in other branches).
  - Constructing EHMM: suppose there are N nodes in the tree, a EHMM state is a size N vector, where each component correspond to a BHMM state. A path from EHMM states gives a multiple alignment (one EHMM state - one column). To construct the EHMM (which states are allowed and how all states should be connected), certain rules must be followed, for example:
    - ∗ Synchronization - the input symbol of a node must match the output symbol of its parent;
    - ∗ State connection must follow that only the active node and its descendents are allowed to change their branch state (active node: the highest-numbered non-wait node).

    See Fig. 2 of [Satija & Hein, Bioinformatics, 2008] for an example of generating alignment by EHMM.
  - Inference: alignment follows standard HMM and multiple alignment - progressive alignment, refined alignment (Virtebi), resampled alignment (Forward); internal nodes either sampled or marginalized.

- Results:
  - Marginalization of internal nodes is better than sampling internal nodes, and resampling generally improves alignment over progressive and refined alignment (Table 1).

- Remark: the composition algorithm is independent of the BHMM model, could use TKF91, or affine gap penalty, etc.

Mouse-human alignment [Mouse genome suppl., Nature, 2002]

- Problems: (for general problem of aligning genomes)
  - Anchor selection: the local alignment procedure for choosing anchors, the statistical cutoff/significance.
  - Chaining of anchors: how to handle rearrangement events.
  - Aligning interanchor regions: microrearrangement events; how to determine which parts are "unalignable".
  - Score model: in both anchor and interanchor alignment steps
  - Handling repeats: transposons as well as SSR

- Methods:
  - Construct syntenic regions: (i) PatternHunter for initial anchoring alignments: high scoring or BBH; (ii) syntenic blocks; (iii) syntenic segments
  - Aligning syntenic regions and fill in gaps: remove or mask repeats by RepeatMasker; 19-base pattern as anchor; extend non-gapped alignment; extend into gapped alignment by dynamic programming; do the 3 steps recursively to fill in the inter-syntenic region gaps.
  - Resolve alignment ambiguity: if one region maps multiple regions in the other species, choose one.
  - Score function: (i) substitution: similar to HKY (strong transition transversion bias) (ii) indel: affine gap penalty.

MCALIGN [Keightley & Johnson, GR, 2004]

- Methods:
  - Sequence evolution model: substitution by JC model, indels by a long-indel model (no multiple-hit correction) where the probability of an indel of certain size is determined from the empirical indel length distribution

- Inference: let S be sequence data, a be alignment, t be time, then sample a by from its posterior distribution: $P(a|S) \propto \int P(S,a|t)P(t)$. Approximate by: $P(a|S) \approx P(a, t_{MLE}(a)|S)$.
- Parameterization: the indel rate parameter θ (relative to the substitution rate) is fixed to be 0.225 (from empirical estimation); and the frequencies of indels are also fit from empirical data.
- Evaluation: compare the predicted alignment and the estimated divergence using the alignment

- Remark: one problem is MCALIGN uses the correct value of θ, which will bias MCALIGN. Use a few different values of θ to show that the performance doesn't crucially depend on θ.

Parametric alignment to Drosophila [Dewey & Pachter, PLoSCB, 2006]

- Methods:
  - For each sequence pair, there may exist multiple optimal alignment (for different parameters). Only when it is a vertex, it is an umbiguous alignment (i.e. always optimal under all parameter settings)
  - Validation: the basic idea is: the functional elements (TFBS) should be conserved.
- Results:
  - Case analysis: Adf1 BS is not conserved in BLASTZ alignment (30% PID), but there are 813 distinct optimal alignments. In paricular, there 2 examples wich 89% and 67% PID respectively.
  - Over all 1,346 TFBS in FlyReg, if fix one set of parameter values, the PID is 79.1%; if for any CRE, choose parameter separately to max PID, then the PID Is 80.4-86.5% (depending on the number of parameters)
- Criticism:
  - Parameters are completely unconstrained, thus ignore the evolution divergence known and the continuity of rates in chromosomes
  - On evaluation: separate alignment (with separate parameters) for each CRE to max PID - not a fair comparison (because you know what to optimize!)

Alignment uncertainty [Wong & Huelsenbeck, Science, 2008]

- Problem: how alignment uncertainty (different programs) affects phylogenetic inference?
- Methods: 7 yeast species, 1502 ORFs, run with different alignment programs
- Results:
  - Phylogenetic tree reconstruction: 46.2% ORFs show multiple trees; furthermore, the tree distance correlates strongly with alignment difference
  - Detecting positive selection: overall estimates of substitution rates are similar; however, in 28.4% ORFs, the inference of positively selected sites is sensentive to alignment methods

2.2 Testing Modes of Selection

Reference: [Yang, Computational Molecular Evolution, Chapter 8; HyPhy manual; Yang, Molecular Evolution: a Statistical Approach, Chapter 11]

Positive selection:

- Selection on protein evolution is defined on $\omega = d_N/d_S$: positive selection if $\omega > 1$; negative selection if $\omega < 1$, and neutral if $\omega = 1$.

- Generally negative selection is the dominant mode in the whole protein, and only a small part of protein may be under positive selection.

Branch model:

- Test: the rates at different branches may be different (but constant across sites), thus testing the difference of rates through LRT. The branches need to be specified *a prior*. Example: ASPM gene (brain size determinant) in human, chimp and orangutan, the rates in three branches are different, and LRT can be used to test if the difference is significant.
- Learning model from the data: when the branches are not specified a prior, could use model selection approach, e.g. AIC. Use the AIC as the measure of fitness of the model, learning the best model through, e.g. genetic algorithm.
- Remark: testing variation of rates in different branches. A strict test of positive selection should require that $\omega > 1$ in the foreground branch.
- Limitation: the selection at the background may be important. Ex. assuming background distribution is uniform (which is actually not): may falsely reject neutral selection on foreground (the model will try to use selection on the foreground to make up for the fact that the background is uniform).

Site model:

- Fixed effect likelihood (FEL): specific sites or regions (from external information) are tested. Ex. in MHC gene, test selection on the ARS (antigen recognition sequence) region: fit with two rates in two regions, and test if $\omega_{ARS} > 1$.
- Site-wise likelihood ratio (SLR) test: for each test, estimate its ω, and test if > 1. Requires a large number of sequences to reach good power.
- Random effect likelihood (REL): the rate of each site is a random variable to be sampled from some distribution. The likelihood of a site h is given by:

  $$f(x_h) = \int_0^\infty f(\omega)P(x_h|\omega)d\omega \approx \sum_{k=1}^{K} p_k P(x_h|\omega_k) \tag{2.3}$$

  where the random effect is approximated by multiple categories of rates (e.g. $K = 10$). Two types of tests that are common:

  - Mixture model test: compare two models: (1) M_0: p_0 fraction sites are under purifying selection $0 < \omega_0 < 1$, and $1 - p_0$ fraction sites are neutral, $\omega_1 = 1$. (2) M_1: p_0 fraction under purifying selection, $\omega_0 < 1$, p_1 neutral $\omega_1 = 1$, and $p_2 = 1 - p_0 - p_1$ under positive selection, $\omega_2 > 1$. M_1 has two more parameters, so use χ^2 test with d.f. 2 (conservative test).
  - Random effect model: compare two models: (1) M_0: the rate follows distribution, $\omega \sim \text{beta}(p, q)$, in the range of $[0, 1]$. (2) M_1: for p_0 fraction sites, $\omega \sim \text{beta}(p, q)$, for the rest, $\omega_s > 1$.
- Identification of sites under positive selection: the rate of any site ω_k for a site k can be computed with the posterior probability. In practice, Empirical Bayesian is used to find the most likely value of ω_k: if MLE of model parameters are used, Naive Empirical Bayes (NEB); or integrate over the model parameters, Bayes Empirical Bayes (BEB).

Branch-site model: [Yang & Nielsen, MBE, 2002; Zhang & Yang, MBE, 2005]

- Test: divide the tree into background (BG) and foreground (FG). The test is whether there are some sites under positive selection in FG.

- Model: test two models M1a and M2a (M1 vs. M2, the suffix 'a' from the modification of an earlier model). Each site belongs to one of four classes:
  - Negative selection (p_0) in both BG and FG: $0 < \omega_0 < 1$.
  - Neutral (p_1) in both BG and FG: $\omega_1 = 1$.
  - Negative selection in BG with $\omega_0 < 1$ and positive selection in FG with $\omega_2 > 1$. The fraction is $(1 - p_0 - p_1)p_0/(p_0 + p_1)$.
  - Neutral in BG with $\omega_1 = 1$ and positive selection in FG with $\omega_2 > 1$. The fraction is $(1 - p_0 - p_1)p_1/(p_0 + p_1)$.

  The null model has $\omega_2 = 1$ (model M1a): so LRT statistic should be compared with a $50 : 50$ mixture of χ^2_1 and 0.
- Other similar models: testing positive selection in clades. Variable substition model: HMM for rates. Switching model: HMM for states (positive selection or not) of AAs over time.

Issues/comments for detecting positive selection:

- The assumptions of uniform d_S: not realistic, should incorporate mutation parameters and possible selection on syn. sites.
- Selection in the codon model: selection ω is uniform (or non-informative prior). Makes biological sense to incorporate chemical properties of AAs. In practice, this is not found to help much, but this could be due to poor correlation of AA chemical properties and selection.
- Detecting one-off directional selection in a particular branch is a main challenge.
- Divergence: if sequences are too similar, not enough statistical signals; if sequences are too divergent, silent changes may have reaches saturation and the data will be too noisy. In simulations, the methods are quite tolerant of high divergence (say with 10 or 50 nucleotide substitutions per nucleotide site along the tree [PAML DOC]), but in practice, high divergence may cause problems in alignment and different codon usage patterns.

Cases of adaptively evolved genes: the most common categories are: immune systems, reproduction, duplicated genes.

Codon substitution models and detecting positive selection [Delport & Seoighe, BriefBioinfo, 2008; Anisimova & Kosiol, MBE, 2008]

- Limitations/directions of codon models and positive selection test:
  - Syn. rate variation: variation of d_S could be caused by site-to-site variation in mutation rates, or by selection on mRNA structure (stability, splicing, etc). This may reduce the power of the test.
  - Empirical codon models: the current, mechanistic, models fail to take into account the selection on AAs, e.g. according to their chemical properties. The combination of empirical and mechanistic models may provide benefits: the fitting of data is significantly better than mechanistic models, but their value in testing selection is not clear.
  - Dependence among sites: the current models assume independence of sites. Possible to incorporate dependence: (1) CpG hypermutation; (2) HMM to capture the auto-correlation of rates in nearby sites; (3) other models of non-local dependence, e.g. Bayesian network model of co-occurring substitutions of pairs of residuals [Poon, PLCB, 2007].
  - Phylogenetic tree: usually a single tree is used, but this may be problematic because of recombination or uncertainty of tree reconstruction. Solutions: (1) Bayesian methods to average over trees; (2) modeling recombination.

  - Alignment and indels: incorporating indels in inference; take alignment uncertainty into account.

- Applications of protein evolution models:
  - Testing mode of selection: functional divergence in different clades (under different selective pressure, thus evolve at different rates); positive selection in certain branches; etc.
  - Identifying functional residuals: if individual residuals are under positive selection, they can be tested for the effect on the protein function.

- References:
  - AA substitution models: [The quest for natural selection in the age of comparative genomics, Heredity, 2007]

Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution [Hubisz & Pollard, COGD, 2014]

- HAR: lineage-specific increase of rate. Could be due to different models: positive selection, relaxation of constraint to GC-biased gene conversion (gBGC).

- Method of detecting HAR: PhastCons to identify conserved elements, then PhyloP to detect lineage-specific change.

- Empirical results: 2700 HARs identified, mean length = 266 bp, often flanked by conserved sequences. Rate in human: 1.7 substitutions per 100 bp vs. 0.2 per 100 bp in conserved sequences and 0.5 in background. Evidence that most of HARs are due to positive selection instead of BGC or relaxed constraint.

- Polymorphism pattern: most HARs are fixed, but rate of polymorphism higher than conserved elements.

- Functional evidence: many are near developmental genes (TFs, genes expressed in CNVs, etc.). 1/3 validated as enhancers in transgenic assay.

2.2.1 Testing Negative Selectoin

SAPF [Sajita & Hein, Bioinformatics, 2008]

- Aim: predict slowly evolving sequences while at the same time accounting for alignment uncertainty.

- Methods:
  - Statistical alignment of multiple sequences: the transducer model and composition from [Holmes, Bioinformatics, 2003]. Use geometric length indel model for BHMM.
  - Phylogenetic footprinting: two sets of states, where one set is fast evolving and the other slow. Output the probability of any position in the reference species being generated from a slow state.
  - Data: Drosophila CRMs: eve 2, eve 3+7 and eme; take FlyReg TFBSs as the gold standard.

- Results:
  - Comparison of SAPF (summing over alignments) vs single best alignment (MPP: maximum posterior product): in 4 Drosophila species (relatively distance), no difference in eve 2, but improves in eve 3+7 (15%) and eme (20%), measured by AUC.
  - Adding species: 3 vs 2 - significant gain; 4 vs 3 - minor gain.

- Remark: summing over alignments is important only when there is considerably alignment uncertainty, e.g. multiple alignment with divergences more than Dmel-Dpse.

2.2.2 Testing Functional Divergence

Testing functional divergence of proteins [Gaucher & Benner, TIBS, 2002; Gu, MBE, 1999, 2001]

- Goal: a protein (usually some residues) may change function after speciation or duplication. Detecting these residues.
- Type I and type II functional divergence:
  - Type I divergence: change of evolutionary rates of proteins, called site-specific rate shifts. Ex. a weakly-constrained residue becomes strongly constrained after it participates to the new function of the protein.
  - Type II divergence: residues are similarly constrained in two clusters, but the residue type and physicochemical property may change across clusters.

Testing type I divergence by correlation: [Gu99]

- Notation: consider two clusters, let T_1, T_2 be the total divergence of two clusters, λ_1, λ_2 be rates in the two clusters (different rates at different residues, thus they are vectors), and X_1, X_2 be the number of AA changes in the two clusters (again per residue, thus they are vectors).
- Rate correlation: if there is no rate change between cluster 1 and 2, then the vector λ_1 should be correlated with λ_2, otherwise, the correlation will be weaker. The fraction of residues that change rates is equal to the correlation coefficient of λ_1 and λ_2. So the degree of type I divergence can be measured by:
$$\theta_\lambda = 1 - r_\lambda = 1 - \frac{\mathrm{Cov}(\lambda_1, \lambda_2)}{\sqrt{\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)}} \tag{2.4}$$
The hypothesis to be tested is thus: $\theta_\lambda = 0$.
- Estimating θ_λ: first compute X_1, X_2 at each residue, and θ_λ can be related to the statistics of X_1 and X_2 (the number of changes is a function of the rate):
$$\theta_\lambda = 1 - \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}} \tag{2.5}$$
where σ_{12} is the covarance between X_1 and X_2, D_1, V_1 (D_2, V_2) are the mean and variance of X_1 (X_2). Note: T_1, T_2 are cancelled out in the calculation.

Testing type I divergence by likelihood: [Gu99, Gu01]

- Simple likelihood method [Gu99]: assume λ is a RV of gamma distribution with parameter α (random variation across sites). The likelihood function:
$$L(\alpha, \theta_\lambda | X) = \prod_k P_k(X_{1k}, X_{2k} | \alpha, \theta_\lambda) \tag{2.6}$$
where k is the site index. The probability term is a mixture of two types of sites: S_0 site - same rate in two clusters, and S_1 site - different rates in two clusters. Note that branch length T_1, T_2 does not enter into calculation, instead, only D_1, D_2 (the mean number of changes) does.
- Likelihood method [Gu01]: the likelihood is computed from the actual AA sequences with the empirical AA substitution model. The parameters are: v - branch lengths, α - shape parameter of the gamma distribution of the rates, θ_{12} - the fraction of sites under type I divergence (i.e. different rates at two clusters). Let X and Y be the AA configuration of some site in the two clusters, respectively, we have:
$$\begin{array}{ll} P(X,Y|S_0) & = \int_0^\infty f(X|\lambda) f(Y|\lambda)\phi(\lambda) d\lambda \\ P(X,Y|S_1) & = \int_0^\infty f(X|\lambda)\phi(\lambda) d\lambda \cdot \int_0^\infty f(Y|\lambda)\phi(\lambda) d\lambda \end{array} \tag{2.7}$$
where S_0 stands for the sites under no divergence and S_1 the sites under type I divergence. The parameters are estimated by MLE, and LRT is used to test the hypothesis $H_0 : \theta_{12} = 0$ vs. $H_A : \theta_{12} > 0$.

- Predicting critical residues [Gu99, Gu01]: by the posterior probability of the site being in S_1:

$$P(S_1|X,Y) = \frac{\theta_{12} P(X,Y|S_1)}{\theta_{12} P(X,Y|S_1) + (1-\theta_{12}) P(X,Y|S_0)} \tag{2.8}$$

Case study: type I functional divergence of caspase family [Wang & Gu, Genetics, 2001]

- Capase family: 42 caspase genes from vertebrates (mammals, chicken, frog), fruit fly and worm. Two broad classes: CED-3 subfamily - apoptotic pathway; ICE family - immune response in mammals/vertebrates. Also different subclasses of CED-3 family may correspond to different death signals: mitochondrial, death receptor (DR), B cell receptor (BCR), etc.

- Type I divergence: comparison of CED-3 cluster and ICE cluster. Estimated $\theta = 0.29$, significant. Using the threshold $P(S_1|X) > 0.6$ gives 21 critical residues with different rates at two clusters (threshold is determined by: removing top 21 residues, θ close to 0). Four residues correspond to positions known important in the functional difference between two clusters.

- Subfamilies of CED-3 and ancestral function: further create clusters within CED-3 subfamily. To study ancestral function, define the distance between two clusters as $d_F(i,j) = \ln(1-\theta_{ij})$, then the distance is additive. Draw the pairwise distance, and the the clusters that remain close to each other may represent ancestral function, and large distance cluster may represent divergent function.

Testing type II divergence: site-specific shift of amino acid physiochemical property [Gu06].

- AA substitution model: define four groups of AAs: charge positive (K, R, and H), charge negative (D and E), hydrophilic (S, T, N, Q, C, G, and P), and hydrophobic (A, I, L, M, F, W, V, and Y). An amino acid substitution is called radical (denoted by R) if it changes from one group to another; otherwise it is called conserved, denoted by C. If no subsitution, denoted by N. The AA substitution model has different rates for three types of changes.

- Model: two substitution models: F_0 - radical changes are not common, reflecting functional constraint; F_1 - radical changes are common, reflecting functional divergence. The evolution of a residue belongs to two cases: (1) no functional divergence: evolution is dominanted by F_0 throughout the tree. (2) Functional divergence: at the early stage, functional divergence (F_1 model); and at the late stage, functional constraint (F_0 model). The ratio of the (2) type of residues, θ_{II} is the parameter to be estimated.

2.3 Evolution of Quantitative Traits

Inferring the historical pattern of evolution [Pagel, Nature, 1999]: once a model of character evolution is available, it can be used for:

- Reconstruction of ancestral states: (i) standard method: generalized least square (GLS); (ii) newer methods: directional GSL method that detects the historical trend [Martins & Hansen, American Naturalist, 1997]

- Estimation of the timing of evolutionary events.

- Tempo/mode of evolution: e.g. clock, gradual change or punctuated;

- Correlation of traits: (i) standard method: independent contrast; (ii) newer methods: model correlated evolution in a likelihood framework (discrete characters); statistical test of the direction of change (temporal-order test) [Pagel, PRSL, 1994]; directional correlation model [Price, PTRSL, 1997]

Brownian motion model of two-node tree:

- Model: given the random variable X_0, the conditional distribution of X_t is given by:

$$X_t|X_0 \sim N(X_0, \sigma^2 t) \tag{2.9}$$

where σ^2 is a rate parameter. Suppose we have a two-node tree with branch length t_1 and t_2 respectively, and the ancestral node $X_0 \sim N(\mu, \sigma^2 t_0)$. We want to know the joint distribution of X_1 and X_2. Note that we could equivalently assume that the root is a constant μ, and X_0 is the least common ancestor of X_1 and X_2, with the distance between the root and X_0 equal to t_0.

- Distribution: first we can show that the joint distribution of X_0, X_1, X_2 is normal:

$$P(X_0, X_1, X_2) = p(x_0)p(x_1|x_0)p(x_2|x_0) \tag{2.10}$$

Since each term is quadratic, the density function is quadratic of x_0, x_1, x_2. Thus the distribution wrt. X_1, X_2 is normal. The expectations:

$$\mathrm{E}\,(X_i) = \mathrm{E}_{X_0}[\mathrm{E}\,(X_i|X_0)] = \mathrm{E}_{X_0}[X_0] = \mu \tag{2.11}$$

where $i = 1, 2$. And the variances according to the law of total variance:

$$\mathrm{Var}\,(X_i) = \mathrm{E}_{X_0}[\mathrm{Var}\,(X_i|X_0)] + \mathrm{Var}_{X_0}[\mathrm{E}\,(X_i|X_0)] = \mathrm{E}_{X_0}[\sigma^2 t_i] + \mathrm{Var}_{X_0}[\sigma^2 t_0] = \sigma^2(t_i + t_0) \tag{2.12}$$

The covariance between X_1 and X_2 according to the law of total covariance:

$$\mathrm{Cov}\,(X_1, X_2) = \mathrm{E}_{X_0}[\mathrm{Cov}\,(X_1, X_2|X_0)] + \mathrm{Cov}_{X_0}[\mathrm{E}\,(X_1|X_0), \mathrm{E}\,(X_2|X_0)] = 0 + \mathrm{Cov}_{X_0}[X_0, X_0] = \sigma^2 t_0 \tag{2.13}$$

Thus (X_1, X_2) follows normal distribution: the mean of X_i equal to the population mean, the variance proportional to the time from the root to the leaf node (i.e. how long the trait has evolved), and the covariance between two leaf nodes proportional to the time from root to the least common ancestor of the two nodes (i.e. how long the two traits have shared evolution).

- The problem of the ancestral node: suppose we have only data of X_1 and X_2, and X_0 and t_0 are unknowns. To make inference of X_0 and t_0, we can see that the MLE of t_0 is always 0, and X_0 a constant μ (if t_0 is not zero, we can see that making it 0 will increase the likelihood). So in the inference problem, we assume that X_0 is a constant, equal to μ.

- Parameter estimation: at $t_0 = 0$, the mean of (X_1, X_2) is (μ, μ), and the variance is $\sigma^2 t_1$ and $\sigma^2 t_2$ respectively, and the covariance is 0. The likelihood function is thus:

$$P(x_1, x_2|\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{(x_1-\mu)^2}{t_1} + \frac{(x_2-\mu)^2}{t_2}\right]\right\} \tag{2.14}$$

The MLE is given by:

$$\hat{\mu} = \frac{t_2}{t_1+t_2}x_1 + \frac{t_1}{t_1+t_2}x_2 \tag{2.15}$$

It could also be written as:

$$\hat{\mu} = \frac{1/t_1}{1/t_1+1/t_2}x_1 + \frac{1/t_2}{1/t_1+1/t_2}x_2 \tag{2.16}$$

Thus $\hat{\mu}$ is the weighted average of x_1 and x_2, with weight proportional to the inverse of the time (inverse of the variance, i.e. the precision).

Brownian motion model of arbitrary tree: [Felsenstein73, Altschul89]

- Model: similar to the model with two nodes, suppose the root is μ, then the leaf nodes follow multivariate normal distribution $N(\mu\vec{1}, \Sigma)$, where $\vec{1}$ is the n-dimensional column vector with each element being 1. The variance of a leaf node is the equal to σ^2 multiplied by the distance from the root to the leaf, and the covariance between two leaf nodes is proportional to the distance from the root to the least common ancestor of the two leaf nodes.

- Parameter estimation: The likelihood function:

$$P(x|\mu) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(x-\mu\vec{1})^T\Sigma^{-1}(x-\mu\vec{1})\right]\right\} \tag{2.17}$$

  Maximize the likelihood function (using the product rule of derivative), we have:

$$\hat{\mu} = \frac{\vec{1}^T\Sigma^{-1}x}{\vec{1}^T\Sigma^{-1}\vec{1}} \tag{2.18}$$

  We can write this as: $\hat{\mu} = \sum_i w_i x_i$, the weighted average of x_i.

Felsenstein's algorithm [Felsenstein73]: solving w_i by matrix computation is inefficient, so we do this by a dynamic programming (DP) algorithm.

- Decomposition of likelihood function: the idea is to factorize the likelihood function in two parts, the first part about some subtree (starting with two closest nodes), and the second part about the rest. Let X_1 and X_2 be two closest nodes, and X_6 be their least common ancestor, we define:

$$u_1 = x_1 - x_2 \tag{2.19}$$

  And

$$x_6 = \frac{t_2}{t_1+t_2}x_1 + \frac{t_1}{t_1+t_2}x_2 \tag{2.20}$$

  Then u_1 is independent of the rest of the tree with mean 0. The likelihood function is now factorized into two terms, the first corresponding to the X_1, X_2 subtree (independent of μ), and the second corresponding to the rest of tree (formed by x_6 and other nodes). Repeat this process, the only term that depends on μ in the end is exponential to $(x_A - \mu)^2$, where x_A is the variable at the root following the above procedure.

- DP for weight computation: the above variable substitution procedure can be written in recursive form. Suppose we have a variable at the i-th node, x_i. Let $L(i)$ and $R(i)$ be the left and right child of i respectively, then we have:

$$x_i = \frac{t_{R(i)}}{t_{L(i)}+t_{R(i)}}x_{L(i)} + \frac{t_{L(i)}}{t_{L(i)}+t_{R(i)}}x_{R(i)} \tag{2.21}$$

  The initial condition is given by the leaf nodes. We could write this in terms of weights, i.e. x_i is represented by a n-dimensional vector, where each component is the weight of the corresponding leaf node in x_i. Thus using weight vector, we have:

$$\vec{w}_i = \frac{t_{R(i)}}{t_{L(i)}+t_{R(i)}}\vec{w}_{L(i)} + \frac{t_{L(i)}}{t_{L(i)}+t_{R(i)}}\vec{w}_{R(i)} \tag{2.22}$$

  And the initial weight is a vector where all elements are zero except 1 (the leaf).

Reference: [Felsenstein, Maximum-likelihood estimation of evolutionary trees from continuous characters, AJHG, 1973], [Altschul & Lipman, Weights for data related by a tree, JMB, 1989]

Weighted average of multiple correlated characters [Vingron & Sibbald, Weighting in sequence space: a comparison of methods in terms of generalized sequences, PNAS, 1993]:

- Problem: given multiple correlated samples, such as sequences or quantitative traits, how should one obtain the average?
  - Generalized sequence/profile: for quantitative traits, average is well-defined. For sequences, need to define average as profiles or generalized sequences, where each position is a multinomial distribution over an alphabet.

  - Intuition: duplicate samples and outliers. The averaging scheme should penalize duplicate samples (each sample adds no new information) or in other words, put more weights on outliers. In general, design the weighting scheme in a way that depends on the distance matrix of samples (thus duplicate samples can be easily identified).

- Idea: the result should be a weighted average of the samples. Let w_i be the weight of the i-th sample, we could define the average of n sample sequences (called S) at the i-th position as:

$$p_w(S_i) = \frac{1}{n} \sum_{k=1}^{n} w_k e_{S_{ik}} \tag{2.23}$$

  where w_k is the weight of the k-th sequence, S_{ik} is the i-th position of the k-th sequence, and $e_{S_{ik}}$ is the unit vector corresponding to S_{ik}. The average could be viewed as the centroid of n points in the sequence space (where distance is defined). Then the desired weighting scheme can be stated in terms of how the centroid should be placed relative to sample points.

- Relationship between weights and the distances to the centroid: suppose w is the weight vector, $D = (d_{ij})_{n\times n}$ is the distance matrix between any two samples, and z is the distance vector of samples to the centroid, i.e. $z_k = d(S_k, p_w(S))$ where S_k is the k-th sample. Then we have:

$$Dw = z \tag{2.24}$$

  We note that $(Dw)_k$ is the average distance of k-th sample to all sample points, thus the relation states that the average distance is equal to the distance to the centroid. The proof is simple (follows from the exchange of summation).

- Two weighting schemes:
  - VA weighting: choose the centroid s.t. it is equi-distant from every sample points, i.e. $z = \vec{1}$ (note that weighting can be arbitrarily scaled). So we have: $Dw = \vec{1}$, and so: $w = D^{-1}\vec{1}$. With this weighting scheme, suppose we have $n-1$ identical sequence and one other sequence (outlier), the centroid would be somewhere in the middle, effectively boosting the weight of outliers.
  - Self-consistent weighting: the distance from the k-th sample to the centroid is proportional to the weight of the k-th sample:

$$Dw = \lambda w \tag{2.25}$$

    So w should be the eigenvector of D.

Phylogenetic average: [Stone & Sidow, Constructing a meaningful evolutionary average at the phylogenetic center of mass, BMC Bioinfo, 2007]

- Motivation: the ACL method [Altschul89] computes average of the tree as the MLE at the root position. However, given a reversible process, the root position is arbitrary, so the average may be biased. This is especially a problem when the tree is very unbalanced (e.g. a lot of branches in one side, and one branch in the other).

- Equivalance of the MLE and the conditional expectation: the ACL method computes the MLE of the root, this is also the expectation of the root conditioned on the observations (the leaf nodes). For instance, for a two branch tree, the MLE of the root is given by Equation 2.16. The conditional distribution of the root:

$$P(x_0|x_1, x_2) \propto P(x_0)N(x_1|x_0, \sigma^2 t_1)N(x_2|x_0, \sigma^2 t_2) \propto P(x_0)N(x_0|\mu_0, \sigma_0^2) \tag{2.26}$$

  where μ_0, σ_0^2 are determined from the qudaratic form of x_0:

$$\mu_0 = \frac{1/t_1}{1/t_1 + 1/t_2} x_1 + \frac{1/t_2}{1/t_1 + 1/t_2} x_2 \qquad \frac{1}{\sigma_0^2} = \frac{1}{\sigma^2}\left(\frac{1}{t_1} + \frac{1}{t_2}\right) \tag{2.27}$$

Thus if we assume the prior distribution of the root is uniform, the conditional distribution of the root is normal with mean equal to the MLE. It can be shown that this conclusion holds for any tree.

- Averaging over the root position: let τ be the position of the root (indexed by the branch and the position in the branch), assumed to be uniformly distribution in the tree. Then the tree average is: $\mu = \mathrm{E}_\tau(\mu_\tau)$ where μ_τ is the average (conditional expectation) when τ is the root of the tree. Alternatively, since:

$$\mu_\tau = w_\tau x \tag{2.28}$$

  We also have in terms of the weight vector:

$$w_{\mathrm{BM}} = \mathrm{E}_\tau(w_\tau) \tag{2.29}$$

  For the two-branch tree, we take expectation of μ_0 in the equation above, over the root position (t_1), and the result is:

$$\mathrm{E}_{t_1}[\mu_0(t_1)] = \frac{1}{T}\int_0^T \left(\frac{T-t_1}{T}x_1 + \frac{t_1}{T}x_2\right) dt_1 = \frac{x_1+x_2}{2} \tag{2.30}$$

- Algorithm for computing the tree average: we denote $w_{k,t}$ the weight vector when the root is at the branch k, $1 \leq k \leq 2n-3$, and the time within that branch t. The computatoin of $w_{k,t}$ follows Felsenstein's algorithm (linear time complexity) We have:

$$\mathrm{E}_\tau[w_\tau] = \frac{1}{T}\sum_{k=1}^{2n-3}\int_0^{L_k} w_{k,t}dt = \frac{1}{T}\sum_{k=1}^{2n-3} t_k w_{k,t_k/2} \tag{2.31}$$

  where t_k is the length of the k-th branch and T is the total length of the tree. The last step is true because of the linearity of $w_{k,t}$ (following from Felsenstein's algorithm). The total time complexity is $O(n^2)$.

- Representative point: the idea is to find a point s.t. the tree average can be obtained by assuming that the tree is rooted at that hypothetic point. This hypothetic point is the representative point of the tree. For the ACL weighting, we have:

$$w = \frac{\Sigma^{-1}\vec{1}}{\vec{1}^T\Sigma^{-1}\vec{1}} \tag{2.32}$$

  It can be shown then that w satisfies the linear equation (using the relation between Σ and D):

$$Dw - z = c\vec{1} \tag{2.33}$$

  where z is the distance vector between the samples and the root. This is true for any root, so taking average over possible roots:

$$Dw_{\mathrm{BM}} - \mathrm{E}_\tau(z_\tau) = c\vec{1} \tag{2.34}$$

  $\mathrm{E}_\tau(z_\tau)$ is the average distance between each species and any point on the tree. Or effectively, we could define the phylogenetic "center of mass", and this is the representative point. Note that: the center of mass is (generally) not in the tree, and instead lies in the extended coordinate space (see Figure 3 of the paper).

- Comparison with VA and ACL weighting: different weighting schemes can be viewed as different choice of the representative point.
  - VA weighting: we know that $Dw = \vec{1}$, thus if we choose $z \propto \vec{1}$, we would have: $Dw - z = c\vec{1}$, so the representative point of the ACL weighting is the point at equal distance to every species.
  - ACL weighting: it has already been shown that the representative point is the root of the tree.

Consider a special case of $n-1$ identical sequence plus one outlier, in BM weighting, the representative point changes as n increases; while VA and ACL weighting do not have this property (Figure 4).

- Remark: the representative point does not lie in the phylogenetic tree, thus has no direct biological interpretation.

Phylogenetic average with graphical models: [Kazemian & Sinha, Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials, PLos Bio, 2010]

- Background: product of normal density. Suppose we have $N(x|\mu_1, \sigma_1^2)$ and $N(x|\mu_2, \sigma_2^2)$, the product of the density functions:

$$N(x|\mu_1, \sigma_1^2)N(x|\mu_2, \sigma_2^2) = N(x|\mu, \sigma^2)N(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2) \tag{2.35}$$

  where:

$$\mu = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}\mu_1 + \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\mu_2 \tag{2.36}$$

  and

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{2.37}$$

  The proof follows from the quadratic form of x. Take the integral of x:

$$\int N(x|\mu_1, \sigma_1^2)N(x|\mu_2, \sigma_2^2)dx = N(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2) \tag{2.38}$$

- Model: we apply the tree graphical model where the variables at the internal nodes are missing. We are interested in finding $\mathrm{E}\,(X_i|O_1)$, the conditional expectation of internal nodes, and the average over i (average over the entire tree, or all internal nodes). The conditional distribution at each branch is specified by Brownian motion model:

$$P(X_i|X_{\pi(i)} = m) = N\left(x_i|m, \sigma^2 t_i\right) \tag{2.39}$$

- Notations: we follow the notations (recurrence variables) of the probabilistic tree model. We can prove by induction that $\beta_i(m)$, $\gamma_i(m)$ and $\alpha_i(m)$ all have normal densities. Thus we only need recurrence in terms of the three variables per node: mean, variance and the constant term. Specifically, for the i-th node, we have:

$$\beta_i(m) = c_i^\beta N(m|\mu_i^\beta, (\sigma_i^\beta)^2) \tag{2.40}$$

$$\gamma_i(m) = c_i^\gamma N(m|\mu_i^\gamma, (\sigma_i^\gamma)^2) \tag{2.41}$$

$$\alpha_i(m) = c_i^\alpha N(m|\mu_i^\alpha, (\sigma_i^\alpha)^2) \tag{2.42}$$

- Upward algorithm: for any internal node i, suppose j and j' are two child nodes of i. We have the recurrence for $\beta_i(m)$:

$$\beta_i(m) = P(O_j|X_i = m)P(O_{j'}|X_i = m) = \gamma_j(m)\gamma_{j'}(m) \tag{2.43}$$

  Using the product of normal density, we have:

$$\beta_i(m) = c_j^\gamma N(m|\mu_j^\gamma, (\sigma_j^\gamma)^2) \cdot c_{j'}^\gamma N(m|\mu_{j'}^\gamma, (\sigma_{j'}^\gamma)^2) = c_i^\beta N(m|\mu_i^\beta, (\sigma_i^\beta)^2) \tag{2.44}$$

  where the constant term, mean and variance are given by:

$$c_i^\beta = c_j^\gamma c_{j'}^\gamma N\left(\mu_j^\gamma|\mu_{j'}^\gamma, (\sigma_j^\gamma)^2 + (\sigma_{j'}^\gamma)^2\right) \tag{2.45}$$

$$\frac{1}{(\sigma_i^\beta)^2} = \frac{1}{(\sigma_j^\gamma)^2} + \frac{1}{(\sigma_{j'}^\gamma)^2} \tag{2.46}$$

$$\mu_{\beta_i} = \frac{(\sigma_i^\beta)^2}{(\sigma_j^\gamma)^2}\mu_j^\gamma + \frac{(\sigma_i^\beta)^2}{(\sigma_{j'}^\gamma)^2}\mu_{j'}^\gamma \tag{2.47}$$

Next we find out the recurrence for $\gamma_i(m)$:

$$\gamma_i(m) = \int_{m'} P(m \to m'|t_i)\beta_i(m')dm' \tag{2.48}$$

Plug in $\beta_i(m') = c_i^\beta N(m'|\mu_i^\beta, (\sigma_i^\beta)^2)$, and integral over m', we have:

$$\gamma_i(m) = c_i^\beta N(m|\mu_i^\beta, (\sigma_i^\beta)^2 + \sigma^2 t_i) \tag{2.49}$$

Thus we have the recurrence:

$$c_i^\gamma = c_i^\beta \qquad \mu_i^\gamma = \mu_i^\beta \qquad (\sigma_i^\gamma)^2 = (\sigma_i^\beta)^2 + \sigma^2 t_i \tag{2.50}$$

- Downward algorithm: initialization, for the root node, $\alpha_1(m) = P(X_1 = m) \sim$ Prior. If i is an internal node, we have:

$$\alpha_i(m) = \int_{m'} \alpha_{\pi(i)}(m')N(m|m', \sigma^2 t_i)\gamma_{\mathrm{Sib}(i)}(m')dm' \tag{2.51}$$

Plug in $\alpha_{\pi(i)}(m') = c_{\pi(i)}^\alpha N(m'|\mu_{\pi(i)}^\alpha, (\sigma_{\pi(i)}^\alpha)^2)$, and $\gamma_{\mathrm{Sib}(i)}(m') = c_{\mathrm{Sib}(i)}^\gamma N(m'|\mu_{\mathrm{Sib}(i)}^\gamma, (\sigma_{\mathrm{Sib}(i)}^\gamma)^2)$:

$$\alpha_i(m) = c_{\pi(i)}^\alpha c_{\mathrm{Sib}(i)}^\gamma N(\mu_{\pi(i)}^\alpha|\mu_{\mathrm{Sib}(i)}^\gamma, (\sigma_{\pi(i)}^\alpha)^2 + (\sigma_{\mathrm{Sib}(i)}^\gamma)^2)N(m|\mu_i^\alpha, (\sigma_i^\alpha)^2) \tag{2.52}$$

where

$$\frac{1}{(\sigma_i^\alpha)^2} = \frac{1}{(\sigma_{\pi(i)}^\alpha)^2} + \frac{1}{(\sigma_{\mathrm{Sib}(i)}^\gamma)^2} \tag{2.53}$$

$$\mu_i^\alpha = \mu_{\pi(i)}^\alpha \frac{\sigma_{\alpha_i}^2}{(\sigma_{\pi(i)}^\alpha)^2} + \mu_{\mathrm{Sib}(i)}^\gamma \frac{\sigma_{\alpha_i}^2}{(\sigma_{\mathrm{Sib}(i)}^\gamma)^2} \tag{2.54}$$

And the constant term:

$$c_i^\alpha = c_{\pi(i)}^\alpha c_{\mathrm{Sib}(i)}^\gamma N(\mu_{\pi(i)}^\alpha|\mu_{\mathrm{Sib}(i)}^\gamma, (\sigma_{\pi(i)}^\alpha)^2 + (\sigma_{\mathrm{Sib}(i)}^\gamma)^2) \tag{2.55}$$

- Average over the tree: first, we know that the conditional distribution $P(X_i = m|O_1) \propto \alpha_i(m)\beta_i(m)$. Plug in the relevant distributions, we have $X_i|O_1$ follows normal distribution $N(m|\mu_i, \sigma_i^2)$ where:

$$\frac{1}{\sigma_i^2} = \frac{1}{(\sigma_i^\alpha)^2} + \frac{1}{(\sigma_i^\beta)^2} \tag{2.56}$$

$$\mu_i = \mu_i^\alpha \frac{\sigma_i^2}{(\sigma_i^\alpha)^2} + \mu_i^\beta \frac{\sigma_i^2}{(\sigma_i^\beta)^2} \tag{2.57}$$

Thus the conditional expectation $\mathrm{E}\left(X_i|O_1\right) = \mu_i$. The tree average:

$$\bar{\mu} = \frac{1}{T}\sum_{i \in T(1)} \frac{\mu_i + \mu_{\pi(i)}}{2} t_i \tag{2.58}$$

- Remark: the variance parameters depend on the Brownian motion rate σ^2, however, if the goal is to estimate the average, σ^2 will be eliminated, and the result does not depend on it. So we could set $\sigma^2 = 1$.

Phylogenetic comparative methods [Harvey & Purvis, Nature, 1991; Martins, TiEE, 2000]

- Problem: find traits whose evolutions are correlated
- Methods:
  - Comparative tests: compare δx and δy in each branch (directional test) or each pair of descendent (undirectional test)
  - Reconstruction of ancestral states
  - Non-parameteric/non-model-based test: pairwise comparison. E.g. if wheneven $x_1 < x_2$ always implies $y_1 < y_2$, then the two traits must be correlated.
- Issues:
  - Ancestral state reconstruction: should include uncertainty
  - Evolutionary maintenance: should also be used in addition to the independent origin [Hansen, Evolution, 1997]
  - Correlation != causation
- Remark: Brownian motion model is flexible: could be interpreted as both mutation-drift and stocastic selection

2.4 Methods of Correlated Evolution

Phylogenetic profiles in prokaryotes and yeast [Pellegrini & Yeates, PNAS, 1999]

- Idea: if 2 genes are functionally linked, then they will share similar phylogenetic profiles; so phylogenetic profiles can be used for predict whether 2 genes are functionally linked.
- Methods:
  - Phylogenetic profile: presence (homolog) or absence pattern
  - Neighbor: based on the distance of phylo. profile (Hamming distance)
- Results:
  - Functionally linked genes tend to have similar phylo. profile (than random gene pairs): comparison of number of neighbors of genes known to linked vs genes randomly chosen
  - Genes with similar phylo. profiles tend to be funnctionally linked. Examples: ribosome, flagellar, histidine biosynthesis.
- Criticism: the accuracy or resolution may be low: for example, most yeast specific genes would have similar profiles (1 in yeast species, 0 in bacterial species), but not all these genes are linked. Similarly, the method cannot distinguish specific aa pathway (his vs trp, etc.)

Phylgenetic profiles in 15 eukaryote species [Barker & Pagel, PLoS Comp. Biol., 2005]

- Methods:
  - Cross-species correlation of phylo. profiles: Fisher's exact test of independence of two discrete RV's (2 RV's can be represented as a contingency table: then apply the usual approach of testing independence for categorical data)
  - Phylogenetic method: via LRT.

        - H1: the two genes are correlated in evolution, a 4-state MC (8 parameters); H0: the two genes are independent (4 parameters).
        - Null distributiion of LRT: since the parameters of H0 is unknown (and different for different pairs), thus sample many random pairs (exclude those with known interactions) and compute the LRT for these pairs

- Results:
    - Improve sensitivity: recover more known interactions than across-species correlation under the same p-value cutoff
    - Improve specificity: across-species correlation predicts more links in random pairs dataset than phylo. method
    - Biological interpretation: why phylogeny is useful?
        - Reduce false positives: ex. two genes may appear to have a high correlation, but in fact this is only caused by a single event, thus insignificant
        - More sensitive: independent gain or loss is important signal for phylo. method
        - The strong predictor is: the number of correlated gain or loss (mainly loss) events
    - Contingent gain or loss of genes: for example gene A is present only if gene B is present, but not the opposite. These relations can be found via the parameters in the MC model.
- Q. parameter estimation of the model: molecular clock assumption (i.e. branch lengths are known)? Estimate for single pairs (15 data points for 8 parameters)? Bayesian?

Coevolution model of RNA interaction [Yeang & Haussler, MBE, 2007]

- Idea: interacting bases (complementary) must change simultanesouly to maintain base pairing. Thus reward the simultanesou change and penalize the individual change.
- Methods:
    - Model: CO model, basic substitution matrix Q of 2 positions is 16 by 16 matrix, assuming the independence of the 2 positions.
    - Parameter estimation: two parameters, $\epsilon < 1$ is the penalty to individual base change, and $r > 0$ is the reward to coordinated base changes. The two parameters are estimated for an RNA sequence using a phylo-HMM model.

2.5 Protein Evolution

Problems: infer the selection force and function of proteins from the molecular pattern of evolution.

- Does the gene change function or not? In paralogs vs. orthologs?
- What domains and residues have more changes, or responsible for functional change? Explanation in terms of structure-function relationship?

Methods:

- Signatures of positive selection: suggest the change of function, and the adaptive value of the change. Intra-species, inter-species (phylogeny), inter- vs intra-species (MK test).
- Signatures of function divergence: type I (rate difference) and type II (conserved but different).
- Remark: the limitation of functional divergence test - the neutral evolution of genes. The protein may change among multiple equally functional configurations (neutral network). The residues are under different constraints in different configurations, thus produce patterns of function divergence.

Hemoglobin: function divergence [Gribaldo & Philippe, MBE, 2003]

1. Background: hemoglobin consists of a tetrmer, 2 identical α subunits and 2 identical β subunits.

2. Methods:

   - Data: 145 α and 145 β from mammals, 46 α and 57 β from teleots (fishes), 45 α and 49 β from Sauropsida (birds, crocodiles, turtles, etc.).
   - Testing site-specific divergence: infer the number of substitutions of each position at each of the three groups (PAML). Then each position is classified as homotachous (constant rate), heterotachous (different rates), constant, and constant but different (CBD).

3. Results:

   - The heterotachous sites: similar in orthologous (mammals vs fish, etc) and in paralogous groups (α vs β), about 30%. About 40 sites found in mammals, but appear evenly distributed all over the structure.
   - CBD sites: overrepresented in paralogous comparison vs orthologous ones, with a mean of 10% and 2% respectively. CBD sites cluster on inter-subunit interfaces.

CFTR: function divergence [Jordan & McCarty, PNAS, 2008]

1. Background: CFTR is a member of ABC superfamily, however, it is a cholide channel. The goal of study is to identify the specific domains and residues most likely to be involved in the evolutionary transition from transporter to channel activity.

2. Methods:

   - Data: CFTR and ABC transporters, 47 paralogous proteins in human, and orthologs in various vertebrate species.
   - Analysis: compare two clusters: CFTR othologs and ABCC4 orthologs. Type I and type II divergence in four domains, respectively, TMD1, ABC1, TMD2, ABC2.

3. Results:

   - Divergence: both type I and type II divergence are significant in both mammals and vertebrates. Type I divergence is stronger, but the difference is smaller in mammals, suggesting further specialization of function in mammals.
   - Specific type II divergence sites: R352 in the sixth transmembrane helix (TM6), conserved in all CFTR orthologs. R352 interacts with D993 (absolutely conserved in all mammals) in TM9 to stabilize the open-channel state. This suggests: CFTR channel activity evolved, at least in part, by converting the conformational changes associated with binding and hydrolysis of ATP, as are found in ABC Transporters, into an open permeation pathway by means of intraprotein interactions that stabilize the open state.

Insuline-like receptor (InR): positive selection at different timescales [Guirao-Rico & Aguade, MBE, 2009]

1. Background: α subunit: extracellular and bind insulin and β suite: trans-membrane and tyrosine kinase activity.

2. Results:

   - Positive selection in two closely related species Dmel and Dsim: MK test. Only detected on the cytosolic domains.

- Evolution of InR in multiple fruit fly species: significant variation in different lineages; significant variation of selective pressure across the gene partitions (Mgene = 0 vs. Mgene = 3); positive selection in the branch connecting Drosophila and Sophophora subgenera - 13 codons in both insulin binding and signaling regions.

A biophysical view of protein evolution [DePristo & Hartl, NRG, 2005]

- Problem: explain the protein evolution pattern using biophysical principles. In particular, reconcile the observation of constant rate of evolution with the observation that most missense mutations have large effects.
- Idea: mutations affect stability of proteins, and the fitness is a function of protein stability.
- Protein biophysics:
  - Improper folding of proteins lead to aggregation and degradation, which are common sources of disease.
  - Protein stability: single-domain and short proteins ($<$ 110 amino acids) follow two-state folding kinetics (folded/native state, unfolded state). The stability can be measured by the free energy difference between the two states, ΔG, typically, -3 to -10 kcal mol^{-1}.
  - Activity-stability trade-off: function of a protein depends critically on mechanical flexibility (functional residues are often destabilizing), thus highly stable proteins tend to have lower activity and difficult to regulate.
  - Effect of mutations: most missense mutations change ΔG by 0.5 to 5 kcal/mol (the same magnitude as ΔG). Almost all mutations, at all sites in a protein, affect stability and aggregation. This is in stark contrast to mutations that affect function, which are generally restricted to a small number of specific catalytic residues.
  - Compensatory mutations: around 10-12 compensatory mutations for each deleterious mutation. In many cases, the primary and compensatory mutations are individually deleterious but jointly neutral.
- Model of protein evolution:
  - Fitness: a function of stability:

$$W(\Delta G) \propto \exp\left(-[\frac{\Delta G - \Delta G_{opt}}{\sigma_{\Delta G}}]^4\right) + c \tag{2.59}$$

  - Distribution of mutational effects: either increase to decrease ΔG, where the offset follows normal distributions for both cases.
  - Evolution: single sequence evolution (if population size is small), and population delocalization (presence of genetic variation).
- Explaining observations:
  - Pathogenic missense mutations are often found to be wild-type in orthologous proteins, knowns as compensated pathogenic deviation (CPD). Two mechanisms (Figure 2) for fixation of a deleterious mutation P: (1) Different protein context: one mutation, C, increases the stability s.t. the mutation P is now neutral. (2) Population delocalization: another mutation C is also deleterious, but the double mutant is neutral. Thus both mutations can be fixed by something like "stochastic tunneling".
  - Molecular clock: the rates of fixation of compensatory mutations are independent of population size.

  - Overdisperson of protein rates (the variance of rates across lineages is large, relative to expection under the neutral model): multiple missense mutations are likely to be fixed simultaneously in large populations, thus increasing the variance.
  - Implications for testing neutrality: with compensatory mutations, one mutation may appear to be under positive selection, whose goal is actuallyto maintain *status quo*.

An overview of protein evolution [Pal & Lercher, NRG, 2006]

- Problems in protein evolution:
  - Integration of various determinants of protein evolution (they are not independent)
  - Fitness & adaptation: what is the fitness distribution of mutations; how adaptation occurs at the molecular level; relative importance of purifying and adaptive selection (more important than what neutral theory asserts).
  - Design principles of protein evolution: robustness (selection favors organisms robust to genetic perturbations); metabolic efficiency of proteins.
- Rate variation of protein evolution: (of the same organism) determined by systematic forces (mutation & recombination) and protein properties.
- Variation of regional genomic properties: Recombination affects the power of selection (selection is less effective in regions of low recombination rate), but its effect is limited (after controlling for coufounding factors such as expression level).
- Purifying selection
  - Fitness density: the fraction of aa that are functional; the selection of overall protein properties such as translational efficiency, mRNA stability, etc.
  - Protein dispensibility: observation - weak correlation with protein evolution. May be caused by: (i) dispensibility may be important to affect evolution rate only after a certain threshold (only when selection is weak); (ii) dispensibility (done in lab) is not a good measure of fitness under natural conditions; (iii). evolution proceeds mainly through point mutations while dispensibility concerns with the influence of the whole protein; (iv) adapation may be important: a newly evoled gene may be environment specific and thus appears more dispensible, will have higher rate.
  - Protein strucure and stability: affects the evolution rate. E.g. the rates of aa at the core and the surface are different. Hypothesis: protein designability (how tolerant of structure to sequence change) variation contributes to the variation of evolutionary rates
  - Pleiotropy: reduces protein evolution. Observation: the pleiotropic proteins (hubs of protein interaction network) evolve slowly. However the effect (if exists) is likely small, after controlling for confounding factors.
  - Expression breadth and expression level: Observation: expression level is the strongest predictor of protein evolution. Hypothesis: proteins with high expression level are less tolerant to weakly deleterious mutations (because these proteins are more likely to form harmful aggregrates if their expression level is high).
- Positive selection:
  - Arms race: between species (e.g. host-pathogen) and within species
  - Compensatory substitutions: e.g. one weakly deleterious mutation may be compensated by multiple positions.
  - Adaptation: lineage-specific changes (e.g. human genes related to brain function)

Functional systhesis approach to protein evolution [Dean & Thornton, NRG, 2007]

- Problem: study sequence-function-phenotype, explain the phenotypic change/difference in terms of sequence variations using functional data about the genes
- evolution of individual proteins for a simple phenotype: insecticide resistence of blowfly.
  - Two phenotypes: sensitive and resistent
  - Identify the aa differences of the candidate gene (known to be acetyle cholinesterase) between 2 phenotypes
  - Mutations that block the activity of the gene through structure and functional analysis (enzyme activity assay)
- Evolution of individual proteins for a quantative/complex trait: mice coat color.
  - Two phenotypes: light and dark (in fact pigmentation level: a complex trait)
  - The aa differences of one candidate gene (many exist) Mc1r
  - Mc1r mutations that change the cAMP signal transduction pathway, and the downstream gene expression
- Evolutionary trajectory: antibiotic resistence
  - Phenotypes: the fitness can be directly measured (or the antibiotic resistence level)
  - Mutations of TEM bete-lactamase: 5 mutations
  - The fitness of each of the 32 possible combinations of mutations
  - Results: only one or two trajectories that are possible (or with signficiant probablity)
- Evolution of pair of interacting genes: Aldosterone-MR (its receptor)
  - Reconstruct ancestral receptor (AncCR): bind both Ald. and Cortisol
  - Duplication of AncCR followed by a mutation in one copy that disrupts the binding to Ald
  - Finally: one copy that is MR and another one that binds on Cortosol (CR)
  - Results: gene duplication provides the material for functional specialization/differentiation; co-option of existing genes (in this case through duplication) for new functions
- Remark:
  - if some trait is correlated monotically with the fitness, then could use this trait to study the evolutionary problems, e.g. the mutational path
  - ancestral sequence reconstruction and if possible the properties of the ancestral sequence: to understand the adaptation

Structure-based protein evolution models [Thorne, COSB, 2007]

- Problem: an explanataory framework for the variation of protein evolutionary rates, e.g. used for phylogenetic inference.
- Variation of evolutionary rates due to mutation:
  - Context-dependent mutations: e.g. CpG mutations - the rate of (methylated) C to T is elevated. The CpG mutation accumulation is dependent on the chronological time (because mutations do not occur in DNA repliacation), wherease other mutations depend on the number of generations.

  Variation of non-synonymous rates among genes:

  - Protein expression level: highly expressed genes are under strong selecton to avoid the translational errors that lead to protein misfolding.
  - Buried vs exposed residues: the buried residues experience replacments more slowly than exposed ones.

  Incorporating structure in evolutionary models:

  - Parisi-Echave study: simulate protein evolution by sequence-sequence compatibility. The simulated sequences have qualitatively similar patterns of variations of preferred residues among protein positions.
  - Statistical models that approximate the Parisi-Echave model without evolutionary dependence.

Structurally constrained protein evolution (SCPE) [Parisi & Echave, MBE, 2001]

- Aim: use the structure conservation to explain why substitution patterns depend on factors such as amino acid physicochemical properties, local structure environment.
- Methods:
  - Structure constraint: the phenotype of a sequence is defined as the distance between its predicted structure and the reference structure. Specifically, the sum of square error of the energy of every position of the protein.
  - Evolution: single sequence evolution under Jukes-Cantor model. A mutated sequence is accepted only if the distance is below some threshold.
  - Test system: The $L\beta H$ domain of the E.coli enzyme LpxA. Multiple alignment of multiple proteins of this family.
- A sigmoidal relationship between sequence divergence and tolerance to structure divergence.
- SCPE accounts for the observed hexaperptide variability pattern: calculate the entropy of each position under simulation (average of multiple runs), and compare the observed entropies.
- SCPE predicts the probability distributions of amino acids at each position of the hexapeptide site. These are compared with the observed patterns.

Positive selection in human-chimp-mouse [Clark & Cargill, Science, 2003]

- Methods:
  - Data: 7645 genes with confident human, mouse and chimp orthologs.
  - Detecting positive selection: H_0: all sites are either neutral ($dN/dS = 1$) or evolve under negative selection ($dN/dS < 1$). The alternative hypothesis: some of the sites are allowed to evolve with $dN > dS$ in the human lineage only.
  - Positive selection at functional categoreis/gene groups: P value distribution of all genes in the group, then Mann-Whitney test for difference beteen the group and the background.
- The neutral null hypothesis was rejected for 28 genes (0.38%) at $P < 0.001$, 178 genes (2.3%) at $P < 0.01$, and 667 genes (8.7%) at $P < 0.05$. This test controls the local substitution rates (through synonmymous substitutions). The rates generally correlate with local GC content, local recombination rate and LINE (repeats).
- The positive selected genes (PSGs) are enriched with genes associated with genetic disease (OMIM).
- Functional categories enriched with PSGs:

  - Olfactory receptors (ORs): It seems likely that the different life-styles of chimps and humans might have led to divergent selection pressure on these receptors. These genes are either undergoing positive selection or are in the process of pseudogenization.
  - AA catabolism: 7 out of 16 genes have $P < 0.05$. Ex. branched-chain amino acid catabolism, which involves the ALDH6A1, BCKDHA, and PCCB genes, is the primary pathway for energy production from muscle protein under starvation conditions.
  - Skeletal development (TLL2, ALPL, BMP4, SDC2, MMP20, and MGP).
  - Neurogenesis (NLGN3, SEMA3B, PLXNC1, NTF3, WNT2, WIF1, EPHB6, NEUROG1, and SIM2).
  - Homeotic transcription factor genes (CDX4, HOXA5, HOXD4, MEOX2, POU2F3, MIXL1, and PHTF), which play key roles in early development.
  - Several genes involved in the development of hearing also appear to have undergone adaptive evolution in the human lineage.

Genome evolution in yeasts [Dujon & Souciet, Nature, 2004]

- Methods: 4 new genomes plus Scer: Cgla, Klac, Dhan and Ylip.
- Protein families: about 40% families (2,000) are common to all 5 yeasts. Most families (1,200) are $1:1:1:1:1$ orthologs.
- Divergence: in terms of AA sequence identity of orthologs: (human-mouse 70%) Scer-Cgla: 65%, Scer-Klac: 60-61%, Scer-Ylip: 48-49%. In paralogs, the divergence is smaller, but show bimodal patterns, some paralogs are highly similar $> 90\%$.
- Lineage-specific gene gains and losses:
  - The most striking example of species-specific gene losses is offered by C. glabrata where 29 genes are lost compared to all other yeasts: (1) galactose metabolism (five genes); (2) phosphate metabolism (four genes); (3) cell rescue, defence and virulence (three genes); and (4) nitrogen and sulphur metabolism (three genes).
  - Specific gene losses were also found in D. hansenii (eight genes missing), K. lactis (five genes missing) and Y. lipolytica (39 genes missing), but their functional coordination is less obvious.
  - A few genes in each species show horizontal gene transfer (ortholog with no other yeast, but with bacterial).

Positive selection in human-chimp [Nielsen & Cargill, PLoS Biol, 2005]

- Background: Human-chimp divergence about 1.5% (substitutions per nucleotide), and 2.3% for CpG islands.
- Methods:
  - Data: 8,079 genes out of which 3,913 were analyzed by [Clark03].
  - Detecting positive selection: accelerated evolution at any point during evolution of humans and chimps (could be either lineages). Classical $dN > dS$ test. Positive selection of gene groups was done by Mann-Whitney test on P value distribution.
  - Polymorphism data for positive selection: top 50 genes on 20 Caucasian Americans and 19 African Americans.
- Functional categories under positive selection:
  - Immune-defense related genes: top of the list. Probably due to the evolutionary arms race between pathogens and host cells.

  - Spermatogenesis: likely causes - sperm competition, selection for reproductive isolation, patholgen-driven selection in the reproductive organs, etc.
  - Olfaction: ORs.
  - Cancer and apoptosis related: may be selective for other functions of these genes. One hypothesis is: selection for avoiding apoptosis during sperm generation (apoposis eliminates up to 75% of sperms, thus any mutation that helps escape apoptosis has a large advantage).
  - Many of PSGs without known functions show sequence similarilty with TFs.

- Expression patterns: only genes in testes show excess of positive selection (each gene is assigned to the tissue where it has maximum expression). Brain seems to be under high level of negative selection.

- X chromosome: positive selection even after removing genes involved in spermatogenesis and expressed in testes. Likely due to the higher selection efficiency of selection at X chromosomes (more effective selection against deleterious recessive mutations and in favor of positive recessive mutations).

- Polymorphism data largerly confirm the positive selection of top 50 genes. In addition, reveal mode of selection: e.g. a developmental gene (SCML1) had 16 fixed substitutions and no polymorphism, suggesting repeated selective fixations, and one OR gene had 6 substitutions and 11 polymorphims, suggesting balancing selection.

- Discussion:
  - Difference with [Clark03]: this study finds targets of positive selection throughout mammalian phylogeny (immunity and defense, spermatogenesis), while [Clark03] is focused on human-specific (OR, etc.).
  - Origin of cancer hypothesis: selection for escaping apoptosis of germ cells, may increase the cancer rates in somatic cells.

Evolutionary principles of gene duplication [Wapinski & Regev, Nature, 2007]

- Problem: what determines the fate of gene duplication? And what are the new functions of duplicated genes?

- Methods:
  1. Data: 17 fungi genome.
  2. Ortholog group identification: SYNERGY algorithm.
  3. Types of ortholog groups: uniform (1 copy per genome) and volatile.

- Volatile grups are enriched with: peripheral transporters, receptors and cell walls, stress response; uniform groups enriched with essential growth process, mitochodrion, ER, etc.

- By transcriptional modules: "cell cycle and meiosis" enriched for uniform and persistent ortho-groups; "development" and "stress and carbohydrate metabolism" are enriched for volatile groups.

- WGD: a different picture, e.g. ribosomal proteins.

- Paralogous genes rarely diverge in biochemical functions (by GO categories), less frequently diverge in cellcular component or molecular interactions (through partners in the physical and genetic networks), and diverge most frequenty at the level of regulation (whether belonging to the same transcriptional modules).

- Discussion:

  - Copy-number variation in stress-responsive genes maynot only be tolerable but beneficial, allowing adaptation to diverse ecological niches. In contrast, genes essential for cell growth, including those necessary for intricate complexes, cannot readily tolerate such noise and tend not to evolve by gradual duplication and loss.

Domain size evolution [Wolf & Panchenko, BMC Evol Biol, 2007]

- Problem: is there any long-term trend of domain size increase or decrease?
- Methods:
  - Data: Conserved Domain Database
  - Detection of the trend of domain size change: correlation of domain size with the distance to the root. If positive, then trend of increase; if negative, decrease.
- A signficant portion of domains show tendency of change: 14-18% increase in size and 5-8% decrease in size
- Similar trend for spacer (inter-domain sequence) size: significant portion of change, and increase is about two times more likely than decrease
- Discussion:
  - Hypothesis: adaptive selection of insertions (novelty)
  - The prevalence of increase is less significant in more recent species. Hypothesis: pressure of small size (for efficiency of transcription and translation)
- Remark: analyze the size evolution from the perspective of protein function. The forces:
  - Adaptive selection of insertions that add novelty
  - Pressure of small size $\Rightarrow$ favor deletions
  - Functional constraint: loop is more tolerable of insertions and deletions

Positive selected genes (PSGs) in six mammalian genomes [Kosiol & Siepel, PG, 2008]

- Methods:
  - Data: only 1:1 orthologs of human and one of the five other species (chimp, macaque, mouse, rat and dog). Quality control: e.g. remove pseudogenes (those that have different gene structure, or low-quality sequencing). Total of 16,529 genes after processing.
  - Method: standard LRTs allowing different evolutionary modes at each site. One LRT on all 6 species (whether there is any selection in the whole tree), and branch- or clade-specific LRTs (e.g. genes under positive selection for some sites in the lineage leading to human). The latter uses the branch-site model of [YN02].
  - Testing enriched GO categories: MW test on P values of all genes in a group.
- 400 PSGs using all 6 species, and 144 branch-/clade-specific PSGs (under FDR < 0.05). Relatively few PSGs in primates: 10 for human branch, 18 for chimp branch, etc.
- Enriched categories in 544 genes:
  - Immunity and defense: particularly enriched in rodent PSGs.
  - Sensory perception: particularly enriched in primate PSGs. OR, five taste receptors (bitter receptor is important to avoid toxic substances).

  - Diet-related genes: MGAM - digestion of starch, MAN2B1 - cleavage of mannose, TCN1 - vitamin B12 transport, steroid hormone metabolism.
  - Other examples: GYPC - regulating mechanical stability of red blood cells, related to malaria susceptibility, CGA - human gylcoprotein hormone (pregency and development, evolution of human endocrine system).

- Gene expression:
  - PSGs overall have lower expression than non-PSGs, and tend to be expressed in more tissues.
  - PSGs in tissues: in the order of PSG enrichment - spleen (immunity), testis (spermatogenesis, liver, breast.
  - Brain genes are not enriched with PSGs and primate PSGs show sharply reduced expression in cerebellum.

Gene gains and losses in Scer genome [Gordon & Wolfe, PG, 2009]

- Methods: reconstruct the ancestral genome (gene order and content) at the point immediately before WGD.
- Gene gains: 124 gene gains, often by duplications (dispersed or tandem), also orphan gene gains. Half of them have no annotation and none is essential when deleted.
  - Alcoholic fermentation: Adh2, Pdc5 and Pdc6 (pyruvate decarboxylase, convert pyruvate to acetaldehyde).
  - Glycolysis and gluconeogenesis: Tdh1
  - Phosphate hydrolysis: Pho3, Pho5 (increase uptake of thiamine, which is a cofactor of PDC and required for alcohol fermentation).
  - Thiamine phosphate (TP): Thi21/22 (biosynthesis), Thi71/72 (thiamine transport).
  - Purine degradation (to eliminate O_2 dependence): Dal4, Dal7.
  - Uptake of sterol under anaerobic conditions: Aus1 and Pdr11.
  - Cell wall remodeling and plasma membrane: Tip1, Tir2/4, Dan2 (in response to cold shock, anaeriobiosis, etc.)
  - Psuedohyphal growth: Ecm23 (a negative regulator, cell wall morphogenesis).
  - Other cases: catabolism of alternative nitrogen sources, drug resistence, defense against oxidative stress.
- Gene loss:: Ura9 replaced by Ura1, decoupling uracil biosynthesis from mitochondrial respiration.
- Discussion: the physiological aspects that adapt to fermentative life style:
  - Carbohydrate metabolism: increasing throughput of glycolysis and fermentation pathways; and inhibiting respiration.
  - Additional metabolism to support fermentation: e.g. thiamine biosynthesis and uptake.
  - Decoupling pathways from dependence on mitochondrial respiration.
  - Reduce the dependence on O_2: bypass those pathways by importing substances from outside the cells.
  - Modification to cell wall and morphology.

Gene Duplication: New Analysis Shows How Extra Copies Split the Work [2016]

- Paper by Xun Lan and Pritchard.

- Question: Why don?t duplicate genes vanish from the gene pool almost as soon as they appear?
- After duplication: more often, the gene duo quickly divvies up the job, e.g. each producing 50% of the same protein level as before. Using GTEx data to show that tissue-expression pattern of duplicates are often similar. Only 15 percent of the time?duplicate genes in humans and mice have clear expression differences in different parts of the body.
- Possible explanation: the duplicates tend to be under the same set of cellular controls. So cis-regulatory changes that weaken the expression compensate for the copy number gain.

2.6 Comparative Genomics of Noncoding Sequences

Dmel vs Dvir comparison [Bergman & Kreitman, GR, 2001]:

- Problem: in noncoding DNA sequences: intergenic and intronic, what are proportion of conserved sequences; pattern of nt. substitutions; pattern of indels?
- Methods: data: 40 loci, 100kb noncoding sequences in Dmel and Dvir (about 40Myr)
- Results:
  - Length distribution of ungapped conserved blocks: similar in integenic and intronic; 22-26% are conserved, with median block length 19bp
  - Substitution pattern of nts in the conserved blocks: transition/transverion bias ≈ 2, lineage specific base compositions.
  - Indel pattern in the conserved blocks: about 20-fold lower than the substitutions; indel length distribution: skewed toward short indels (1-5bp) with a long tail of long indels. The mean and median are 7.73bp and 2bp respectively

Dmel, Dere, Dpse, Dwil and Dlit comparison [Bergman & Celniker, GB, 2002]:

- Methods:
  - data: 5 Drosophila species, covering 500kb sequences of Dmel genome
  - CNCS (conserved noncoding sequences): identified in VISTA using a window size of 10bp with PID $\geq 90\%$
- Results:
  - microsynteny: largely conserved at the scale of individual fosmids (40kb) in Drosophila. Found only 6 rearrangments. Not conserved with Anopheles: widely scattered. Ex. 27 genes in Dmel Chr. 2R are distribution in Anapheles genome chr. 2L (10), 2R, 3L and 3R.
  - CNCS clustering: spacer length between CNCS: non-random distribution (reject the exponential distribution predicted by mutation cold spot hypothesis). Homologous spacer lengths are highly correlated between Dmel and Dpse (R = 0.85): suggest functional constraint of spacer sequences
  - CNCS predicts enhancers: overlap with known enhancers; predict a new enhancer in the ap region: verified by experiment
- Criticism: the correlation of the lengths of the homologous spacers cannot lead to the functional constraint. Intuitively, longer spacers will also be longer in the orthologous sequence due to phylogenetic inertia.

Ubiquitous constraints on non-coding sequences in Drosophila [Halligan & Keightley, GR, 2004; 2006]:

- Background:

  - Drosophila genomes are very compact: high deletion rate in Drosophila; very few pseudogenes (about 100 found)
  - Genes of complex expression pattern are correlated with large intergenic size

- Methods:
  - Orthologous sequence pairs: orthology: through BBH of exons; intronic sequence; intergenic sequence: boundary defined via start/stop codon of adjacent genes (equally divided in the two genes)
  - Alignment: initial alignment: genome alignment tool MAVID; realignment on 500bp anchor-defined regions (anchor: ¿10bp ungapped matches): MCALIGN2
  - Choosing neutral sequences: (1) FEI: fastest evolving intronic sites, 8-30 bp region of introns; (2) FEF: fastest evolving four-fold degenerate sites, in genes without codon bias, in the center (less constrained than edge)
  - Measuring constraint on pairwise alignment: 1 - O/E where O = observed number of of substitutions; E = expected number of of substitutions of neutral sequences

- Result:
  - Divergence of sequence is negatively correlated with intron size: short introns (<60-80 bp): high divergence; large introns: low divergence. Hyp: large introns contain functional sequences (short introns are devoid of them, close to mimimal size allowed)
  - Divergence of sequence is negatively correlated with intergenic size (but weaker than intron size)
  - Constraint pattern in introns: very high in the edge, due to splice sequences; large introns are more constrained
  - Constraint pattern in intergenic sequences: large intergenic sequences are more constrained
  - Constraint estimation on the whole genome:
    - introns <= 80bp: very small constraint (0.18-0.20)
    - introns > 80bp, intergenic sequences (both 5' and 3'): large constraint(>0.5)
    - Hyp: a) unannotated transcribed sequenes (RNA or protein-coding); b) CRE
  - Clustering of substitutions: variation of the distribution of functional sequences: comparison with geomestric distribution of distance of adjacent substitions.

- Remark: the constraint pattern of sequences can suggest functions: level of constraints, types of sequences under more constraint, spatial distribution of constraints, etc.

Adaptive evolution of non-coding DNA [Andolfatto, Nature, 2005]:

- Problem: the constraint of non-coding DNA? Adapative evolution?

- Data source: X-linked Dmel genes, 12 Dmel alleles and a single Dsim sequence, 31 coding and 51 non-coding regions (27 IGR, 24 UTR)

- Result: let IGR be inter-genic region
  - Non-coding DNA has a lower divergence than the synonymous mutations: estimated constraint is 40% in introns, 60% in UTR and 50% in IGR. Test: pairwise divergence corrected for multiple-hits
  - Non-coding DNA has a lower polymorphism than the synonymous mutations. Test: compare the polymorphism between the two groups by Wilcoxon test; Tajima's D test: the distribution of rare alleles

  - Adaptive selection in non-coding DNA: 60% in UTR and 20% in introns and IGR. Test: (1)McDonald-Kreitman test of divergence vs polymorphism (not signficant in IGR and introns). (2) Modified McDonald-Kreitman test: remove mutations that are rare (these are sites under negative selection) to increase the sensitivity of the test

Positive selection in promoters of human genes [Haygood & Wray, NG, 2007]:

- Summary: prediction of promoters with positive selection. Validation/functional characterization of these promoters with gene functional category, the gene expression patterns and changes.
- Methods:
  - Data: 6,280 human promoters (upstream 5K), alignment with chimp and macaque.
  - Detecting positive selection: null model - only neutral and negative selection (mixture model); alternative model - neutral, negative selection and positive selection. Also, in order to distinguish positive selection from relaxation of negative selection, allow the fraction of negative selection changes in human lineage. Test the two models with LRT. Controls: introns of nearby sequences.
- Results:
  - Promoters under positive selection: 46 with Q value < 0.05, and 575 extra with $P < 0.05$, corresponding to $Q = 0.55$ (extra 250 promoters).
  - Functional classes of significant promoters: neural and nutrition (glycolysis, carbonhydrate metabolism) related are most significant categories.
  - Promoters under positive selection vs. tissue specificity of expression (some tissue specific genes may be under more positive selection, e.g. metabolism related): one significant correlation - pancreas.
  - Expression divergence and positive selection on promoters: negative correlation (expected), though not significant.

Human-specific gain of a developmental enhancer [Prabhakar & Noonan, Science, 2008]:

- Methods:
  - Detection of human accelerated conserved noncoding sequences (HACNS): testing via likelihood-ratio of human lineage specific changes.
  - Functional assay: transgenic mouse enhancer assay, using lacZ reporter gene.
- Results:
  - HACNS1: expect 4 substitutions in human, while 16 were seen, highly significant. 13 are clustered into a 81 bp window. Predict gains of BSs of PAX9 (known expression in the limb) and ZNF423 (also known expression).
  - Expression pattern of HACNS1: strong limb expression domain relative to chimp, possibly adaptive (human hand dexterity).

Arabidopsis intragenomic conserved noncoding sequence, [Thomas & Freeling, PNAS, 2007], Conserved noncoding sequences (CNSs) in higher plants [Freeling & Subramaniam, COPB, 2009]:

- Genomes of plants are more diverse than vertebrate genomes: increased duplication events, polyploidy, increased recombination, transposable elements (TEs) and gene silencing [Reineke, NAR, 2011].
- Plant CNSs average from 20-30 bp in length, and the most conserved plant CNSs are much less conserved than are the most conserved animal CNSs.

- Intragenomic comparison: from more than 3,000 gene pairs, identify 14,944 intragenomic Arabidopsis CNSs. The mean CNS length is 31 bp, ranging from 15 to 285 bp. Each gene is on average associated with 1.7 CNS.
- Lessons for plant CNS prediction: [Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes, Reineke & Gu, NAR, 2011]
  - Clear differences in upstream region evolution between monocots and dicots, suggesting a separation of these groups should be made.
  - A divergence time of about 100 mya sets a limit for reliable conserved non-coding sequence (CNS) detection.

Evolutionary analysis of regulatory sequences (EARS) in plants [Picot & Ott, Plant J, 2010]:

- Challenges:
  - The recent duplication events undergone by many plant genomes is a major complication, as many of the duplicated genes have acquired new regulatory features associated with altered functions
  - Far fewer sequenced plant genomes.
- Data: 2000 bp upstream of translational start sites as putative promoter sequences (average intergenic region for Arabidopsis, may be much larger in other species, e.g. rice).
- Pairwise comparison: alignment of 2000 bp sequences, and convert the alignment into p-values. Note that the comparison strongly depend on the species compared. Ex. many peaks above threshold when comparing with Brassica (closely related); but few peaks pass the threshold when comparing with rice.
- Multi-species conservation profile: geometric mean of pairwise p-values.

2.7 Genome Organization & Evolution

1. Genome plasticity [Watanabe & Gojobori, JME, 1997]

   Methods:

   (a) Data: Haemophilus influenzae, Mycoplasma genitalium, Escherichia coli, and Bacillus subtilis

   (b) Methods: identify clusters of genes in conserved order (allow gene insertions and deletions) in pair-wise comparison.

   Results:

   (a) Dynamic rearrangements are very frequent, even breaking the operon structure (not directly discussed).

   (b) Identified some clusters that are highly conserved, suggesting strong structural constraints. The longest cluster is comprised of three operons, the S10, spc, and a operons. Collectively, they encode 26 ribosomal proteins, the RNA polymerase a subunit, and the preprotein translocase SecY subunit.

2. Instability of operon structure [Itoh & Gojobori, MBE, 1999]

   Methods:

   (a) Data: Ecoli and BS operons, in 11 genomes.

   (b) Classification of operon conservation: Identical, Similar (allow translocation and gene insertion/deletion), Destructed (genes present, but separated) and Unknown (no gene).

Results:

(a) In general, operon structures are not very conserved, and the degree of conservation seems to be correlated with the distance of species.

3. Infer functional coupling from gene clusters [Overbeek & Maltsev, PNAS, 1999]

   Methods:

   (a) PCBBH or PCH: a pair of genes X and Y is PCBBH (pair of close bidirectional best hits) or PCH (pair of close homologs) if their distance in two genomes are smaller than a gap size threshold

   (b) Scoring of gene pairs: score of a PCBBH or PCH is equal to the divergence between 2 species; and the score of a gene pair is the sum over all PCBBH or PCH where the pair occurs

   (c) Parameters: gap size = 300bp, determined from mean and standard deviation of distance of related genes; and the same strand

   (d) Gene clusters: from repeatedly merging related gene pairs

   Results:

   (a) Reconstruction of purine biosynthesis pathway: the different gene presence profile in Gram positive and Gram negative species. The unexpected gene in the cluster: yexA homology (hypothetic cytosolic protein).

   (b) Reconstruction of glycolysis pathway.

   (c) Partition the pairwise connection network into 343 clusters.

4. Uber-operons [Lathe & Bork, TIBS, 2000]

   Methods:

   (a) Gene neighborhood: similar to [Overbeek, PNAS, 1999], if two genes are within 250bp, and in the same strand.

   (b) Uber-operons: repeatedly merge conserved gene neighbors.

   Conclusion: although genomic rearrangements cause variation in the immediate neighborhood of a gene, many genes are maintained over evolutionary time within the context of a discrete set of functionally related genes, called "uber-operons". A possible scenario is a type of purifying selection.

5. Genome alignment [Wolf & Koonin, GR, 2001]

   Methods:

   (a) Genome (local) alignments: produce a set of conserved gene strings, which are gene clusters with conserved order.

   (b) Pairwise comparison only.

   Results:

   (a) Coverage of conserved gene strings of the genome: generally very small, thus suggesting a high rate of gene order change. Implications of genome evolution: the intra-genome rearrangment events are more important in shaping the genome than horizontal gene transfer and lineage-specific gene loss.

   (b) The direction of genes in the same conserved gene string: mostly in the same direction.

(c) Functional predictions: only consider those clusters which appear in at least 3 genomes. Ignore the functional links between genes with well-known functions. The function of an unknown COG is assigned by the cluster it appears (majority voting?) as well as sequence information (in some cases, e.g. the domain of this gene) ⇒ a table of predicted function of 90 COGs with the species they occur.

6. Conservation of gene pairs [Ermolaeva & Salzberg, NAR, 2001]

   Problem: score gene pairs (the probability of being in the same operon) by their distance in multiple genomes

   Methods:

   (a) Score gene pairs: must be in the same direction and close (< 200 bp) in both genomes. Determined by the number of operons and number of directons (consecutive genes in the same direction) in both genomes; and the distance of the species (smaller if they are close)

   Results:

   (a) Analysis on 34 microbial genomes and find 7600 pairs that are highly likely to be operons ($> 98\%$)

7. SNAP [Kolesov & Frishman, JMB, 2001]

   Motivation: construct the functional gene groups without requiring colinearity

   Methods:

   (a) SN-graph (similarity-neighborhood graph): a gene pair is linked in this graph if they are in the same strand and close in multiple genomes

   (b) SN-cycle: the closed path in SN-graph

   (c) Measuring functional coupling of genes in the same SN-cycle: (i) KEGG pathway annotation; (ii) enrichment of MIPS category

   Results:

   (a) The statistical significance of SN-cycles: long SN-cylces disappear in the shuffled genomes when the phylogenetic distance is large. Therefore, most of the long SN-cycles should be significant.

   (b) The predictive accuracy of SN-cycles: predict the function of one gene in a SN-cycle.

   (c) Application: annotation of genes of a new prok. genome.

8. Gene order conservation in Prok. [Tamames, GB, 2001]

   Background:

   (a) Gene orders are highly conserved in close species, but reshuffled significantly in distant species.

   (b) There are still highly conserved gene clusteres in very distant species, suggesting selection.

   Methods:

   (a) Data: about 35 genomes.

   (b) Conserved gene clusters: called "runs", clusters of genes in which order is conserved. A run cannot comprise genes from different strands; hence a change of coding strand implies the termination of the run. Allow the maximum of 3 genes inserted in gaps.

   (c) Measuring gene order conservation between genomes: the number of genes belonging to conserved runs divided by the total number of shared genes.

   (d) Measuring conservation of gene clusters: (Fig. 4) the number of genomes where the cluster is conserved (genes that belong to the conserved runs). Drawn in a plot that display the distance.

Results:

(a) The extent of gene order conservation between genomes is negatively correlated with the phylogenetic distance.

(b) The most conserved runs: (using E.coli. as reference) (i) with some exceptions, every run is preferentially composed of ORFs belonging to the same functional class; (ii) usually correspond to operons in E. coli, and combinations of two or even three operons are common, suggesting that additional factors, other than common regulation, acting in the conservation of gene order.

Remark: the evolutonary constraint is measured essentially by the number of genomes, or considering distance only qualitatively, and no statistical testing.

9. Connected gene neighborhood [Rogozin & Koonin, NAR, 2002]

Motivation: find "uber-operons": the gene cluster in each genome is only part of a super-operon.

Methods:

(a) Conserved gene pair := gene pairs that are in the same direction and separate by ≤ 2 genes in at least 3 (out of 31) genomes

(b) Connected gene neighborhood (gene arrays): connect the conserve pairs (find the "tiling path" through the conserved gene paris)

Results: find 188 clusters of gene arrays: most of them have coherent functional themes

10. Functional modules [Snel & Bork, PNAS, 2002]

Aim: identification and characterization of modules from the pair-wise constructed network.

Methods:

(a) Data: 34 prok. genomes.

(b) Network construction: associate two orthologous groups if they co-occur in the same potential operon two or more times.

(c) Network partition: identification of linkers and partition into clusters

(d) Functional homogenity of clusters: each orthologous group is defined by its COG category, then the homogenity of a cluster is the entropy of the COG categories. The statistical significance is assessed by random sampling of clusters of the same size.

Results:

(a) Functional analysis of clusters: (i) 70% have a more homogeneous functional composition than that of a random cluster of the same size; (ii) 50% of the within cluster enzyme pairs belong to the same pathway vs 9% of between-cluster pairs.

(b) Function prediction of genes: e.g. an uncharacterize gene in flaellum cluster.

(c) Function of linker proteins: tend to have duplications, which lead to different functional associations and thus different clusters.

11. Gene Teams [Luc & Raffinot, CBC, 2003]

Results:

(a) Predicted gene teams in three genomes, mostly 2 or 3 genes.

(b) Trp operon: all genes are present in the 3 genomes, but the order is not conserved (Fig. 4).

12. Gene neighborhood analysis: review [Rogozin & Koonin, Briefings in Bioinfo., 2004]

Results:

(a) Concepts: (i) functionally related vs evolutionarily conserved gene clusters: the latter is significantly only when the species are divergent; (ii) computational difficulty of gene order analysis in comparison with DNA or protein sequence analysis.

(b) Methods for identifying conserved neighborhood: dot plot, alignment, conserved pairs, colinearity-free approach, etc.

(c) Applications: function prediction, cis- element analysis, phylogenetic reconstruction, etc.

13. Comparative genome structure in prok. [Bentley & Parkhill, ARG, 2004]

    Results:

    (a) Gene orientation: there is a bias from 52-83%, depending on the genomes. Geenomes with high bias tend to encode a protein involved in DNA replication.

    (b) Conservation of gene order: (i) although there seems to be a positive selection for clustering of physically interacting proteins, there is no absolute requirement for juxtaposition of any genes in a bacterial genome and synteny is lost at a much faster rate than sequence similarity (e.g. two genomes may share many genes, i.e. high sequence similarity, but the conserved gene strings are infrequent); (ii) inversions seem to be centered on the origin of replication.

14. Predicting functional modules [Wu & Xu, NAR, 2005]

    Methods:

    (a) Functional relation between a gene pair: based on three sources of information. (i) phylogenetic profile: the co-evolution of two genes; (ii) gene neighborhood: the score is determined by the distance of the two genes in all genomes; (iii) similarity of GO assignments

    (b) Functional modules: from "segmentation" of the graph constructed from pairwise relations

    Results: find 185 modules in E. coli genome

15. Operon prediction by DVDA [Edwards & Wernisch, NAR, 2005]

    Method: directon vs directon analysis

    (a) Compute the probability that two genes are from the same operon if they belong to the same directon in both genomes

16. Nebulon [Janga & Moreno-Hagelsieb, NAR, 2005]

    Methods: for each gene, find its related genes. A related gene is defined by the number of links of the two genes - if two genes appear in a single operon, then they are linked.

    Results:

    (a) Validation of the method: use KEGG to validate the predicted links.

    (b) Nebulon recovers some uber-operons.

17. GeneChords: phylogenetic method for conserved gene clusters [Zheng & Kasif, BMC Bioinformatics, 2005]

    Aim: detecting conserved gene clusters while taking into account of the phylogeny.

    Methods:

    (a) Conservation of gene pairs: two-state process (1: presence; 0: absence), and the measure is effectively BLS.

(b) Merging gene pairs: each gene is assigned a upstream conservation and downstream conservation scores, based on pairwise values above. Then choose the boundary of cluster by these two scores (upstream boundary: high downstream score, but low upstream score; etc.). The score cutoff is determined by reshuffling the genome.

Remark: a typo in the paper, should be: $-\log P(X = 1|Y = 1) \propto d(X, Y)$ (negative sign), where Y is the immediate ancestor of X.

18. Relaxed gene teams [Kim & Yang, CSB, 2005; Kim & Yang, JBCB, 2006]

Motivation: operons are often split in different genomes, thus want to relax the proximity constraint when doing multiple-genome comparison

Methods:

(a) Program parameters: T_z: minimum gene set size; T_δ: proximity threshold; T_p: minimum number of genomes where the proximity constraint is satisifed; T_s: minimum number of genomes where the gene cluster appears as a whole. The program computes all gene clusters for any given set of parameter values

(b) A gene cluster satisifing all constraints is called a hybrid gene pattern.

Results:

(a) Analysis of 97 genomes: running time; the dependency of number of clusters on the parameters

(b) Operon prediction in B. subtilis: (i) 48 experimentally verified operons are recovered, out of which 26 have exactly the same boundaries; (ii) those outside the boundaries, or those not overlapping with operons are functionally related (through literature)

(c) Phylogeny reconstruction using common gene clusters

(d) Genome annotation: annotate a gene by its neighbors in multiple genomes (voting)

Remark: 20345 total clusters were predicted with 97 genomes. 26 out of 48 known B. subtilis operons are recovered in this set. But this set is so large that it does not say anything about the functional coherency/accuracy of this set (many of them could be false positives, i.e. from the shared ancestral gene order).

19. Domain teams [Pasek & Raffinot, GR, 2005]

Aim: the conserved cluster of domains.

Methods:

(a) Generalized gene team: out of a given set of genomes, if a gene team occurs in more than one genome, then it will be in the result. Therefore, the number of teams can be exponential.

(b) DomainTeams algorithm: similar to gene teams except that domains are basic units instead of genes. The algorithm is exponential to the number of genomes.

Results:

(a) The identified domain teams can detect rearragement events such as domain fusions. An example domain team from 5 bacteria in (Fig. 4).

(b) Identify all domain teams in 15 Gram negative bacteria and compare the result with 309 E.coli operons: 245 were fully recovered. The phylogenetic distribution of the 245 operons (for each operon, the number of species that it occurs).

Remark: the value of gene teams formulation lies in the ability to detect clusters with rearrangement, which will be difficult to find under the iterative merging scheme.

20. Life cycle of operons [Price & Alm, PLoS Genetics, 2006]

    Conclusion: operon evolution, including its formation, is driven by selection on gene expression patterns. Changes in operon structure are shown to be associated with changes in gene expression patterns, so the diversity in operon structure may reflect adaptation to differing lifestyles.

21. GCQuery [Yang & Sze, GR, 2008]

    Goal: for a given gene cluster (query), identify its counterparts in other genomes to study its evolution.

    Methods:

    (a) Significant orthologous cluster: extract the orthologs of the query genes, if they are close, then signficant. Tested by the hypergeometric test: the number of query genes occuring in a window of given size. The probability has to be modified by E-value: under the null hypothesis, how many clusters will be found in a window.

    (b) Definition of gene cluster (order) conservation: the percent of neighboring gene pairs that are conserved (not divided by break points).

    (c) Data: 123 E.coli operons; 400 bacterial genomes in 18 groups; BLAST for identifying homologs.

    Results:

    (a) Operon occurrence: wide distribution (Fig. 4), only 3 operons are conserved across all 400 genomes.

    (b) Gene orientation: in 92% operons, the gene orientations are fully conserved.

    (c) Operon rearrangement: most clusters, 85%, have perfectly conserve neighboring gene pairs.

2.8 Genome Comparison: Case Studies

A high-resolution map of human evolutionary constraint using 29mammals [Nature, 2011]

- Data: 29 mammals, total divergence 4.5 substitution per site, comparing with 0.68 for human-mouse-rat-dog. Increasing div. is important, e.g. at single base, $p < 0.02$ to be perfectly conserved, and $p < 10^{-25}$ for 12-bp element.

- Detecting constrained elements: SiPhy-omega. Found 3.6M constrained NCEs, 4-5% of genome, and median length 36 bp (much shorter than previous efforts).

- Annotating constrained elements: (1) Synonymous constrained elements; (2) conserved RNA secondary structure: e.g. Xist. Typically short elements, e.g. hairpin (20 bp or less); (3) Promoters: enriched GO terms in constrained vs. less constrained promoters; (4) Conserved motif instances: overlap strongly with ChIP-seq TFBS; (5) Chromatin states: an example of SNP in Hox enhancer, overlapping a FOXO motif. Overall, 30% constrained NCEs coding, 30% chromatin states, 1.5% RNA, 3% conserved motif instances.

- Detecting positive selection at individual codons: Stepwise Likelihood ratio (SLR) method (1) estimate codon model at gene level: equilibrium codon frequency, transition/transversion ratio, dN/dS of gene; (2) estimate dN/dS for each site, fixing other parameters, with LRT. Found about 2% of codons under positive selection.

- Detecting HARs: first identify about 1M conserved NCEs (excluding human), then use PhyloP to test accelerated evolution in human. Found about 500 HARs.

- Q: In SiPhy method, how to obtain neutral rates? Similar to PhyloHMM, which uses mixture model to obtain neutral model?

- Q: the UCSC PhastCons scores, what is the resolution?

Genome sequencing reveals insights into physiology and longevity of the naked mole rats (NMR) [Kim & Gladyshev, Nature, 2011]

- Background: NMR biology. Longevity and cancer resistance: challenge of the theory of redox homeostasis (non increase of cancer at old age). Live in full darkness and low oxygen. Divergence to rats: 70M years.
- Testing positive selection on genes: use PRANK to align the sequences (aware of phylogeny); filtering of gaps. Branch-site test from PAML.
- Results: possible genome adaptations
  - Telemere pathway: positive selection and unique AA changes.
  - Visual perception and olfactory: gene gain and loss.
  - Cancer resistance: transcriptome change (TGF-beta/SMAD3 over-expression in old NMR).
  - Thermoregulation: unique AA change in UCP1.
  - Adaptation to low oxygen: change of transcriptome.
- Lesson: to use genome data to reveal genetic basis of adaptation, we do (1) gene level changes: expansion or gene loss. (2) Positive selection. (3) Unique AA. (4) Transcriptome comparison.

Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the Arctic [Lynch & Schuster, Cell Reports, 2015]

- Divergence: about 5Mya, relatively few changes fixed: about 1-2 per gene.
- Detecting possible positive selection: in general, most fixed substitutions are probably neutral, according to the Neutral Theory. However, in this case, there is likely very strong selection, so the fixed substitutions are likely enriched with positive selection. Find gene groups enriched with fixed substitutions.
- Adaptation in: body plan (e.g. tail length, coat color). Temperature sensing (TRPV3). Circadian biology.

Chapter 3

Regulatory Evolution

3.1 Overview of Regulatory Evolution

Problems of regulatory evolution:

- Goal: let G be genotype, P be phenotype (expression) and F be fitness, then evolutionary process can be expressed as: $G \rightarrow P \rightarrow F$. The goal is to dissect the evolutionary changes within a system (ΔG, ΔP, ΔF), and how they are related.
- Are phenotypic variations genetic, and to what extent they are due to non-genetic/environmental forces?
- Selective forces on the regulatory system: are the evolution of regulation under selection? What mode of selection?
- Mechanisms of regulatory changes: what are the immediate causes of the changes: the contribution of trans- and cis- changes? How the overall regulatory network evolves to accompolish the changes? In particular, how are the changes coordinated (overcome the fitness barriers)?
- Mechanisms of conservation: e.g. if expression does not change, whether cis-regulatory sequences are also conserved? If not, how expression is conserved despite sequence changes?
- What does the evolutionary pattern tell about the underlying regulatory systems? Ex. constraints on enhancers, on gene expression/co-expression.
- Relation to phenotypic evolution: how the changes of regulatory systems are relaed to the evolution of phenotypes.
- Remark: often study the evolution at multiple levels, e.g. genome level and epigenome/transcriptome level. This is important for understanding both the immediate causes (e.g. sequence changes lead to transcriptome change) and ultimate causes (e.g. whether TFBS changes lead to change of expression) of the observed evolutionary patterns.

Specific questions and hypothesis of regulatory evolution:

- Expression evolution: to what extent the variation across species is due to genetic (sequence level) changes?
- Consequences of regulatory changes: are they neutral or functionally important? E.g. do divergent TF binding profiles correspond to functional changes?
- Divergence of CREs and gene expression: sources of adaptation (positive selection)?

- Multi-level nature of regulatory evolution: changes at a lower level may not affect the higher level, e.g. TFBSs gain and loss but not affect gene expression.
- What contribute to the evolutinary changes of gene expression (or other traits)? cis- vs. trans? Distributed (many small changes) or concentrated?
- Inferring the mechanisms/constraints of enhancer sequences: e.g. enhancersome vs billboard model; and regulatory networks.

Evolution of transcriptional regulation [Wray, MBE, 2003]

- PRINCIPLE: the patterns of CRM evolution, including compositional patterns such as TFBS gain, loss and spatial patterns such as conservation of space between adjacent TFBS, are determined by the functional properties of CRM (how senstivie the function of a CRM is to spacing, etc.), thus, the pattern can be used to infer information about CRM.
- Remark: the pattern must be interpreted in the context of various evolutionary forces, not just selection. This is particularly important when selection is weak.
- Discussion: various hypothesis has been made regarding the pattern of CRM evolution
  - Spatial pattern:
    - Core TFBS or neighboring TFBS: spacing tends to be under selection
    - The order and spacing of TFBS: in general not very conserved
  - Heterogenity of evolution:
    - Importance of TFBS → gain/loss and substitution rate of TFBS
    - Specificity of TFBS → evolutionary rate, e.g. high specificity → low rate
    - The function of CRM → evolutionary rate as a whole. Ex. control module is under stronger selection than booster modules, proximal slower than distal, etc.

Variable gene expression in eukaryotes: a network perspective [Wittkopp, J Exp Biol, 2007]

- Expression variation:
  - Intra-species variation: expression difference among individuals underlies the phenotypic difference. Ex. the differential response of reproductive, worker and soldier ants to JH. Technique: mapping of eQTL
  - Inter-species variation: expression divergence between species underlies the phenotypic divergence
- Expression variation and gene regulatory networks: how the role of a gene in the GRN determines the evolution of its expression.
  - Hyp1: GRN influences the distribution of mutational variance of gene expression. Ex: terminal genes have higher mutational variance because of larger trans-mutation target; the genes under negative feedback control have smaller mutational variance.
  - Hyp2: GRN influences the natural selection of gene expression. Ex: the top-level genes are under stronger selection than terminal ones; the genes under negative feedback control evolve more slowly.
  - Hyp3: functional class of genes influences the evolution of gene expression.

Evolution of transcriptional control in mammals [Wilson & Odom, COGD, 2009]

- Expression evolution in mammalian tissues:

  - Examples of cis-regulatory evolution: pigmentation of stickleback fish, malaris resistence in primates.
  - Conservation of tissue-specific expression: even without sequence conservation [Chan & Hughes, JB, 2009]
  - Divergence of expression in primates: may contribute to species-specific traits.
- The TF landscape:
  - About 1,300 TFs annoated, most are either tissue-specific or generally expressed.
  - TF binding specificities of 100 DBDs: highly conserved, half of them have a secondary motif preference.
- Divergence of TF binding and gene expression:
  - Examples of TF binding divergence: liver tissue, stem cells and E2F2 binding in multiple tissues.
  - TF binding divergence often have little impact on gene expression.
  - Gene expression divergence between human and mouse: hybrid mouse with chr. 21, TF binding is almost identical, thus expression change is mainly contributed by the change of CRS [Wilson & Odom, Science, 2008]

3.1.1 Evolution of cis-Regulation

Evolution of CRMs [Ludwig, COGD, 2002]

- Models of CRM evolution:
  - Stabilizing selection: a large number of sites contribute to the fitness of CRM s.t. the average selection for each site is small ⇒ allow relatively high rates of fixation of nearly neutral mutations
  - Compensatory selection: within CRM
- Implications of the two models:
  - Neutral evolution of nonfunctional sequences, e.g. spacers
  - Neutral or nearly neutral evolution of TFBS due to the ”fuziness” of motifs
  - Evolutionary accretion of new TFBSs s.t. each one becomes marginally important
  - Co-evolutionary changes among multiple BS s.t. no net functional change of CRM

Structure-function-evolution of CRS in Drosophila [Wittkopp, Heredity, 2006]

- Conserved sequence, conserved function:
  - Many evidences of conserved enhancers with conserved function
  - Counter-evidence: Dmel-Dpse comparison only reveals a small fraction of known enhancer sequences [Richards & Gibbs, GR, 2005]
- Divergent sequence, conserved function
  - Evidence: (i) eve stripe 2 CRM [Ludwig & Kreitman, Development, 1998; Ludwig & Kreitman, Nature, 2000; Ludwig & Kreitman, PLoS Biology, 2005]; (ii) polymorphism of phylogenetically conserved TFBS [Balhoff & Wray, PNAS, 2005]
  - Explanations of conserved function despite sequence divergence: degeneracy of TF binding; redundent TFBS and redundent enhancers; change of TF without changing the output; coevloution of TF and its binding sites.

- Divergent function: change at CRS
  - de novo evolution of TFBS
  - Duplication and divergence of enhancers (on duplicated genes)
  - Modification of existing enhancers [Wittkopp & Carroll, Curr Biol, 2002; Gompel & Carroll, Nature, 2005]
- Discussion: sequence conservation is insufficient to detect enhancers?
  - Explanation: the data of enhancer sequence divergence concerns the change of TFBS, but the enhancer sequence overall is still more conserved
  - Evidence: [Li & Halfon, GB, 2007; Papatsenko & Dubchak, Genomics, 2006]

Cis-regulatory mutation and phenotypic evolution [Wray, NRG, 2007]

- Methods: comparison within population or between closely-related species → identify the cis-regulatory differences associated with change of traits
  - Within population: genotype-phenotype association stuides using SNP
  - Between species: the CRE sequence under positive selection
- Examples: in human
  - Malaria resistence: Gene: DARC (IL8R); Mechanism: -46 mutation of GATA1 BS in DARC → abolish expression of DARC in red blood cells (but not other types) → malaria resistence
  - Lactase persistence: Gene: lactase (LCT), the enzyme that digests lactose; Mechanism: 4 independent mutations in CRE of LCT (at MCM6 intron) → induce expression of LCT
  - Cognition:Gene: PYDN, neural peptide, related to cognition; Mechanism: 5 mutations in intron + 1 DREAM BS → different of PYDN expression in the brain

Rewiring of TRNs [Tuch & Johnson, Science, 2008]

- Problem: rewiring of TRNs, i.e. switching of transcriptional regulators, are often observed. Why?
- Hypothesis: cooperative interactions among TFs (generally, weak and not very specific) facilitate TRN changes. Expression of a gene depends on regulators, sequences, and cooperative interactions between TFs, thus the change of one can be compensated by another.
- Evidences:
  - Correlation between number of regulators and the fuzziness of the CRSs [Bilu & Barkai, GB, 2005].
  - Simulation of systems with interacting components: the existence of redundant intermediate states, greatly catalyzed changes within systems [Haag & Molla, Evol Int J Org Evol, 2005].
  - Evolution of PPIs: example in yeast a-sepcific genes (Mcm1-$\alpha 2$ interaction).

3.1.2 Evolution of Transcriptome

Cross-species expression comparison to identify genes important for phenotypes

- Idea: the cross-species expression comparison can be used to identify candiate genes underlying some traits, just as cross-species sequence analysis is used to find functional sequence elments.

- If a gene is involved in a conserved process (e.g. kernal of a developmental GRN), then the expression of this gene should be under stabilizing selection ⇒ find candidate genes through conservation of expression pattern

- If a gene is involved in an adpative process (e.g. terminal differentiation of a developmental GRN), then the change of expression of this gene should be correlated with the change of the corresponding trait ⇒ find candidate genes through evolutionary correlation of expression pattern with traits.

- Remark:
  - Strategy 1. is most appropriate for comparison of distant species: most genes have diverged thus those whose expression profiles are conserved are important
  - Strategy 2. is most appropriate for comparison of close species: most genes have not changed much thus those whose expression profiles have changed (in particular correlate with the change of trait) are important.

Adaptation/Adaptive change of gene expression:

- Difficulty: distinguish random drift and adaptive change.

- Idea: if the change is associated with some other change, then it cannot be explained by chance alone
  - Association with phenotypic change
  - Coevolution of two or more genes or modules: correlated change in multiple genes; coevolution of two genes across multiple species; coordinated change of modules. Ex. a module of 10 genes, 9 has increased expression in one vs another species; while the number of increase expected under neutral model for this module is 5. Thus the change of this module is likely to be adaptive.
  - Association with cis-regulatory sequence change

Models of expression evolution:

- Strategy: treat gene expression as quantative trait, and use the testing of quantative traits to test the evolutionary mode of gene expression.
  - Discrite evolution model: hidden variables (activated or inactivated)
  - Continuous evolution model: Brownian motion model; Khaitovich-Pabbo model.

Common issues for cross-species comparison of expression:

- Remove experimental noise for conservation analysis. Methods: use ANOVA to estimate the mean expression of a gene in each species (per condition), then the hypothesis is formulated in terms of the evolution of the mean expression.

- Remove experimental noise for adapative analysis. Methods: use ANOVA to estimate the mean expression of a gene in each species (per condition), remove those genes which fail ANOVA null hypothesis (no significant change), and use the estimated mean for evolutionary analysis

- How to make sure that genes are comparable. Methods: (i) normalization of expression level of each gene. Ex. finite-mixture model clustering. (ii) LRT for each gene which takes into account of the specific expression level and variance of that gene

- Use of population varianance in the Brownian motion model. Motivation: in different experimental conditions, the variances of the same gene could be different (or the variance depends on the expression level), want to model this variance in the evolutionary model

- Testing correlation between continuous variables; between one continuous and one categorical.

- If a gene has multiple measurements in different conditions or tissues: how to treat the profile?

Natural selection on gene expression [Gilad & Rifkin, TiG, 2006]

- Theoretical basis of testing selection: different mode of selection predicts different rate of interspecies divergence vs intra-species variation (mutational variance):
  - Stabilizing selection: few genetic variations are fixed across species, thus divergence $<$ neutral estimate based on intra-species variation
  - Adapative selection: more genetic variations are fixed across species, thus divergence $>$ neutral estimate based on intra-species variation
- Patterns of selection on quantitative traits: influence both mean and variance of the traits
  - Strong purifying selection: the same mean and low variance both between and across species.
  - Weak purifying selection: the same mean across species but larger variance within species.
  - Directional selection: difference in the mean.
- Testing selection by comparing with the neutral model:
  - Prediction under the neutral model: if there is no non-genetic component to expression variation, the increase in the variance between populations (fixed differences) will be proportional to the within population variance (segregating changes).
  - Neutral divergence: can be predicted based on levels of segregating genetic variation, the effective population size and the number of generations separating populations. THus comparing with neutral divergence can tell if a gene is under stabilizing or directional selection.
  - Challenges in practice: partition genetic and non-genetic variance is hard for non-model organisms.
- Observation: stabilizing selection is the dominant mode of gene expression evolution, while directional and neutral evolution play smaller but important roles.

Evaluating the role of natural selection in the evolution of gene regulation [Fay & Wittkopp, Heredity, 2007]

- Problem: change of gene expression due to adaptation or mutation-drift?
- Correlation of expression change with some trait in a phylogenetic tree: if a gene whose expression change correlates with the trait change in the phylo. tree, then the change is likely adaptative and this gene is likely a candidate gene for this trait.
- Methods: phylogenetic comparative method.
- Issues:
  - Brownian motion model fails to capture stabilizing selection
  - Intra-specific variation: cannot use comparative method because of the shared history
  - Genes identified in this way may not have a causal role in the trait change: it may be the effect of trait change, rather than the cause
  - The dependence among coregulated genes is not modeled
- Test of neutrality based on the rate of expression change: neutral model of quantative trait evolution and testing against the neutral expectation.
  - Genetic model: multiple loci, each making small effect
  - Mutation model: effect of mutation could be discrete or continuous (Poisson or Gaussian)

  - Random genetic drift

- Tests:
  - Intra- vs inter- species pattern of variation
  - Only rank genes without estimating the mutational variance
  - Relative rate test: compare rates across lineages (rate of Brownian motion)

- Issues:
  - Error from expression measurement
  - Constant mutational variance is not a valid assumption

- Empirical pattern/rate of expression change
  - The rate of change of neutral (pesudo) genes
  - Nutation line (all mutations are accumulated): empirical measure of mutational variance

Comparative biology of gene expression [Tirosh & Barkai, COBT, 2007]

- Characteristics of gene expression evolution
  - Expressions are condition dependent: must be accounted for, i.e. the inter-species comparison should be done in the same conditions
  - Gene modules: genes work as groups
  - Coexpression: one could compare the expression relative to other genes - coexpression

- Types of comparison:
  - Single gene, single feature (condition)
  - Single gene, multiple features
  - Coexpression of two or multiple genes
  - Gene modules

- Similarity/Conservation of gene expression:
  - Observation: in most available dataset, the conditions in different species are not directly comparable
  - Idea: conservation of coexpression

- Difference/Adaptation of gene expression: distinguish the changes due to random drift and those due to adaptation.
  - Coherent differences in large sets of functionally related genes [Ihmels & Barkai, PLoS Genetics, 2005]
  - Associated changes at the cis-regulatory level [Ihmels & Barkai, Science, 2005]

Comparative studies of gene expression and the evolution of gene regulation [Gallego Romero & Gilad, NRG, 2012]

- Infering modes of selection from expression comparison:
  - Theoretical work: derive neutral divergence from mutational input, generations and so on, then compare the observed with neutral expected divergence. Hard to do for non-model organisms.

  - Empirical approach: rank genes according to pattern of expression [Gilad & White, Nature, 2006]. (Box 2)
  - Interpreting patterns: for directional changes, may not be due to positive selection, e.g. from environmental changes.

- Results about pattern of selection in primates:
  - Negative selection: majority of genes evolve under selective constraint. Empirical, and theoretical estimate (conservative estimation).
  - Directional selection: 10-30% of genes under positive selection. Ex. a gene involved in Vitmain A mechanism, probably related to diet change.
  - Tissue specificity: different patterns of selection in different tissues. A large-scale study identifies both highly conserved tissue-specific expression module; and modules with lineage-specific expression patterns.

- Molecular mechanisms of expression evolution:
  - TF binding divergence in mammals: however, hard to evaluate the extent that binding divergence account for expression divergence.
  - Histone modification: large difference across three mammals, but highly conserved near TSS. Explain 7% of expression divergence.
  - DNA methylation in promoters: may account for 12-18% of expression level difference.

- Relation to phenotypic evolution: e.g. deletion of 5kb sequence upstream of androgen receptor gene in human, likely related to human-specific loss of sensory birissae and penile spine.

3.1.3 Evolution of Regulatory Proteins

TF evolution [Wagner & Lynch, TIEE, 2008]

- Evidence of cis-regulatory evolution:
  - Christmas tree model: regulatory landscape defined by TF gradients, and CREs interpret the regulatory landscape. No pleiotropy problem.
  - Evidence of Christmas tree model: wing and abdominal pigmentation pattern, birstle pattern in fruit flies.
  - Criptic homology: convergent evolution of traits, best explained by the cis-evolution model.

- Evidence of TF evolution :
  - Sequence conservation pattern of TFs: islands of conservation in a sea of divergence. Ex. Hox genes. And there is evidence that the divergent domains are functional - adaptive selection, etc.
  - Many AA differenence between Hox homeordomains are involved in PPIs, and some differences are associated with phenotypic changes.
  - Orthologous TFs may not be functionally equivalent: e.g. tinman (fly) and Nkx2.5 (vertebrate) - both in heart development.
  - Paralogous TFs may have different functions: e.g. HoxA-11 and HoxA-13.

- Escapeing negative pleiotropy of TF evolution:
  - Alternative splicing: tissue-specific alternative splicing increase protein diversity and may help escape negative pleiotropy. Ex. AML1. Specific-specific exons are common.

  - Short linear motifs (SLiMs): often mediate PPIs, just 3-10 AA long with only 2-3 AAs absolutely required for PPI. Tend to be poorly conserved. Ex. Ftz: function in segmentation in fly but as homeotic genes in beetle and grasshopper, the function change is associated with the loss of YPWM motif in fly lineage.
  - Simple sequence repeats (SSRs): abundant in proteins regulating gene expression and evolve rapidly. Ex. SSR variation of Alx-4 is associated with extra claws in different dog breeds, probably through change of interaction with LEF-1.

- Model: TFs generally have modular structure, capable of interacting with different proteins (at different tissues) at different parts of the protein sequence. Thus changing one part of sequence may change interaction with one co-factor and affect a subset of target genes, without pleiotropic consequences.

Lineage-specific TFs and GRN evolution [Nowick & Stubbs, BIFG, 2010]

- Gene duplication and expansion of TF families:
  - Three types of duplications: WGD (twice in vertebrate evolution), segmental duplication (SD) and retrotransposition. SD are often generated in tandem.
  - Duplication as sources of novelty: Most gene duplicates become pseudogenes within a few million years. The survived genes often follow two paths: subfunctionalization and neofunctionalization. As a result, they are often under positive selection, have different expression patterns. Examples: in human, immune defense, reproduction, intercellular communication, neutral development and transcription are enriched in familes tha have expanded through SD.
  - Expansion of TF famillies: e.g. glucocorticoid receptor (GR) expansion in worms, Zn-Cys6 doamins are fungi specific, KRAB-ZNF genes show large expansion in human-chimp. The new TFs may form different partners (or acqure different other domains) and change functions. Ex. C2H2-ZNF domain pairing with a KRAB domain dominates in vertebrate genomes.

- Structure and evolution of GRN:
  - Developmental GRNs: kernels, terminal selector genes/motifs.
  - Dynamics of GRN in different conditions: TFs can fulfill different functions depending on the the conditions and that TFs that serve as hubs in one condition might not be hubs in another.
  - Kernels are highly conserved, but may serve as drivers of species differences. TFs with high connectivity: evolutionary patterns are ambiguous (under more constraint or not?).
  - Co-expression networks: many co-expressed gene pairs are not conserved between human and mouse.

- Cases of GRN evolution and phenotypic changes:
  - The duplication of Runt and acquisition of new network partners early during vertebrate evolution opened the door for development of teeth and skeleton in fish and tetrapods.
  - Change in BMP4 expression is the main determinants of the beak morphology in Darwin finches.
  - Human and mouse stem cells: WNT pathway and BMP in pluripotency in mouse but differentiation in human.
  - Cell types in the brain: quantitative continuous variations like the ratio of co-expressed genes.

3.2 Evolution of Posttranscriptional and Posttranslational Regulation

The evolution of N6-Methyladenosine in primates [Ma & White, 2015]

- Pattern of M6A sites across human-chimp and rhesus: about 13K M6A sites in each species, or about 2 sites per mRNA. Overlap between two humans: 70% peaks are shared; across species: human-chimp, 63%, human-rhesus, 50% and all three 40%.
- Conseration and lineage-specific changes of M6A: use mixed model to defined conserved or human-gain, human-loss sites, where species effect is fixed, and individual effect random. Total of 2861 conserved and 320 Human Gain, and 30 Human Loss events.
- Evolution of M6A site sequences:
  - Motif occurrences: highest in conserved sites, but lower in human-loss sites.
  - Ka/Ks test of M6A sequences: compare the divergence of M6A motif and other sequences in the peaks. The ratio is 0.55 in conserved sites and 3.75 in human-gain sites (consistent with negative and positive selection).
  - MK test: The NI (Neutrality Index) in "human gain" and conserved m6A peaks are 0.58 and 2.2.
- Correlation of M6A evolution with expression: positive correlation between m6A modification and the abundance of m6A-modified mRNA.
  - Expression variation: highest in lineage-specific M6A sites, and lower in conserved sites.
  - Human-gain M6A sites show higher expression than conserved ones than human-loss M6A sites.
  - Within human genome, highly expressed genes have higher m6a modification.
- GO and KEGG analysis of M6A conserved genes and lineage-specific genes: conserved genes are enriched with housekeeping function, while human-gain show enrichment of regulation and DNA binding.

3.3 Regulatory Evolution in Yeasts

1. Evolution of regulatory networks of metabolism in fungi [Lavoie & Whiteway, COM, 2009]

   Cases:

   - Lipid biogenesis.
   - Glycolysis/galactolysis.
   - Ribosome biogenesis.
   - AA biosynthesis/tRNA-aminoacylation: Gcn4 regulation is largely conserved, but how Gcn4 is activated is different in some lineages such as Calb (dependence on Gnc2 is weak).

2. Fungal regulatory evolution [Anne Thompson & Regev, FEBSL, 2009]

   Goal: expression changes within or across species, their adapative roles and mechanisms of changes (cis- and trans-).

   Signature of expression evolution:

   - Data: both intra-species (Kruglyak group, Hartl group) and inter-species (Gasch group, Barkai group).
   - Observation: low ED (expression divergence) genes tend to be associated with growth control and metabolism and high ED ones with respones to internal and external signals (e.g. stress response). Evidence from isogenic cells, intra-species and inter-species variations.

   Promoter changes and expression divergence:

   - TFBS turnover is common and may not lead to expression divergence.

- Promoters of low ED genes tend to have: well-positioned nucleosomes, most of CRS are located with nucleosome free region, transcription if TATA-independent and less susceptible to chromatin remodeling.

Changes in trans- factors:

- Changes in binding specificity of TFs: Rpn4 [Gasch04].
- Changes in TF targets (even if CRS does not canges): Ste12 and Tec1 [Borneman, Science07]. Low-affinity binding may be important.
- Duplication of TFs, other changes such as dimerization, interacting with other TFs.

Contribution of cis- vs trans- factors:

- Intra-species variation: about 70% of intra-species variation can be attributable to trans-effects.
- Inter-species variation: cis-effects seem to dominate from the inter-species hybrid studies.
- Model: trans- variation is more subject to dominance effects than cis variation. Thus on shorter timescales, the pervasive pleiotropic effects and the much higher rate by which trans variation is produced can account for the gene expression variation observed within populations. Over longer timescales, purifying selection could purge trans-regulatory variation. Conversely, although cis variation is produced at a slower rate, positive selection may act more efficiently to fix cis changes due to their higher additivity and weaker pleiotropic effects.

Adaptation and empirical studies of adaptation:

- Few known examples: MRP expression changes in Scer and relatives involve loss of RGE; variation in sporulation efficiency in different Scer strains can be mapped to TFs, Ime1, Rme1 and Rsf1 [Gerke & Cohen, Science09].
- Experimental evolution: adapation to nutrient limitation (glucose, sufate or phosphate). Multiple changes may lead to the same solution, e.g. increase of glucose transporter can be achieved by changing the transporter itself (amplification) or the regulator of the transporter [Gresham, PG08].

3. Population variation of expression and phenotypes in yeast [Fay & Eisen, GB, 2004]

   Problem: genes associated with the copper sulfate resistence in yeast

   Methods:

   (a) phenotype: CuSO4 resistence; rust coloration
   (b) expression data: 9 yesat strains; cDNA array where reference = equal sample of RNA from each strain; measure gene expression under 2 conditions: YPD (rich medium) and CuSO4
   (c) candidate gene identification: a gene is a candidate gene if its expression difference in the two conditions is correlated with the phenotype Phenotype is encoded as a binary variable (resistence or not).
   (d) testing of adaptive evolution: correlation of expression difference and sequence divergence (suppose sequence evolution follows a molecular clock)

   Results:

   (a) phenotypic variation: 2 strains are sensitive to CuSO4 (M34 and YPS163) - reduced growth rate
   (b) correlation between phenotypic and expression variation: choose r ¿ 0.8 ⇒ identify candidate genes (Figure 4)
   (c) the overall expression difference follows a clock (i.e. correlates with the sequence divergence), but not the expression of the candidate genes ⇒ suggest these genes are under (+) selection

4. Parallel inactivation of GAL pathway [Hittinger & Carroll, PNAS, 2004]

   Background: different ecology of yeasts. Ex. galactose is widely used, while some substrates such as the plant-made fructose polymer inulin, are used by fewer species. In particular, S. kudriavzevii (Skud) were isolated on decaying leaves or soild unlike its relatives in sugar-rich environment.

   Methods:

   (a) Data: genome sequences of Scer, Spar, Smik, Skud, Saby, Scas, Cgla, Sklu, Kwal, Klac and Egos.

   Results:

   (a) Parallel loss of galactose utilization and GAL pathway: Skud, Cgla, Kwal and Egos cannot utilize galactose, and GAL pathway is lost for all of them except: (1) Gal4 is retained in Egos (presumbly it regulates other functions); (2) in Skud, pseudogenes of GAL (multiple stop codons and frameshifts).
   (b) GAL pathway loss in Skud: the promoter sequences show different patterns in different GAL genes:
       - Gal80 and Gal4 promoters are more similar to other species than the promoters of structure genes in GAL pathway, suggesting different selective constraints during pathway degeneration.
       - Gal4 promoter retains the Mig1 binding site (glucose repression).
       - Non-GAL promoters that are targeted by Gal4: Gcy1, Mth1 and Pcl10 retrain Gal4 binding site.
   (c) Order of GAL pathway inactivation: no evidence of ordered inactivation, from the number of substitutions.

   Remark:

   - Pleiotropy is one of the main constraints of evolution: Gal4 is also a regulator of other genes, thus under stronger constraint. And since Gal80 is a regulator of Gal4, it may be also under stronger constraint.
   - Pathway gain/loss may follow certain order: (not obvious for pathway loss) e.g. even if galactose is not needed, if Gal4 is inactivated first, the other genes may be mis-regulated (and create waste).

5. Conservation and difference of GRN in fungi [Gasch & Eisen, PLoS Biology, 2004]

   Problem: is GRN (regulatory relations) conserved or changed in different species of fungi?

   Methods:

   (a) Constructe motif-gene group mapping: cluster genes and learn motifs from each group. 35 unique gene groups in S. cerevisiae (Sc) and 42 motifs found (incluidng the ones learned by MEME).
   (b) Gene groups examined: cell cycle and mating type; metabolism (amino acid biosynthesis, purine/pyramidine biosynthesis, ribosomal proteins, phosphate metabolism, etc.); stress response.

   Results:

   (a) Conservation of cis-regulatory system: Fig. 2.
       - Orthologous gene groups often use the same motif (50-75% motif-group mapping are conserved across very large distance, where most non-coding sequences cannot be aligned)
       - The extent of motif-gene group mapping conservation correlates with divergence
   (b) For the orthologous gene groups where Sc motifs are not enriched, novel motifs are found(Fig 3). The correctness of motifs are verified via conservation in multiple species divergent enough. This suggests that the orthologous gene groups may still be co-regulated, but are under the control of different factors.

(c) Conservation of spacing of motifs: (i) bias of motif positions in multiple species; (ii) cases where the distacne between 2 interacting motifs is smaller in target groups than genome-wide control. Therefore, motif positions may be under constraint (still considerable change of site positions, Figure 4).

(d) Case study: proteomoe gene group show change of motif in Sc and C. albicans (Ca). The motif of this group is Rpn4p, which has different PWMs in two species; furthermore, show experimentally that Sc-Rpn4p and Ca-Rpn4p have different binding specifities. It was found that: Sc-Rpn4p can bind strongly with Sc sequences and Ca-Rpn4p can bind with Ca sequences.

Discussion:

(a) Conservation of cis-regulatory networks: overall highly conserved. However, the extent of conservation does not correlate with the importance of gene groups. For example, ribosomal proteins and proteasome proteins are all highly conserved genes (in terms of protein and expression), but their cis-regulatory mechanisms are not.

(b) Models of GRN evolutions:

  - Addition of new gene targets into an existing network: evolution of new CREs. Ex. genes related to buding in Sc are also expressed in G1, and they share the same motifs in other G1 genes (MCB, SCB).
  - Co-evolution of TF protein sequence and CREs: e.g. proteasome genes where Rpn4p sequence and its binding sites co-evolve, while maintaining binding.
  - GRE rewiring: different TFs may regulate a gene group. Ex. all cases where a different motif can be identified in the orthologous gene group.

Remark:

(a) To fully address the problem of GRN conservation and change, one needs to know GRN. To overcome this lack of knowledge, assume there are "hidden" TFs controlling groups of gene whose motifs can be learned through gene clusters. From this view, GRN can be represented as a mapping from motifs (possibly many) and gene groups.

6. Evolution of ribosomal regulation in yeast [Tanay & Shamir, PNAS, 2005]

   Problem: the change of CRS of genes/modules with conserved expression pattern.

   Methods:

   (a) Data: Scer and Spombe expression compendium; 17 yeast genome data

   (b) Conserved transcriptional modules: use SAMBA algorithm (biclustering) to find txp. modules in two species, and take the intersection. Then project to each of the 17 species (Projected Orthologous Modules, POM)

   (c) Motif finding in each of the POM in all species: used to analyze the change (gain/loss) of motifs

   Results:

   (a) Conserved txp. modules and the motifs: some conserved modules also share sequence motifs; but other ones do not. Example: ribosomal proteins (RP) modules, stress response modules

   (b) Analysis of motifs in RP modules: Homol-D motif in RP genes in some species, but Rap1 motif in other species; and yet other species contain both sites. Hyp: ancestor has Homol-D BS, but during evolution, switch to Rap1 binding in some lineages (including Scer).

   (c) Further evidence of regulatory switching of Homol-D to Rap1:

      - In the species containing Rap1 sites, found the Rap1 gene with the txp. activation (TA) domain, i.e. the appearance of TA domain exactly matches the appearance of Rap1 binding sites (in the same branch of the phylogenetic tree).

- For the “intermediate” species, evidence of interaction between Homol-D and Rap1: the distance between the two binding sites in most genes are very small (about 6bp).

Discussion/Remark:

(a) Model: during evolution, Rap1 acquires the TA domain, and works with Homol-D in the target genes of Homol-D (perhaps PPI); later Rap1 completely replaced the role of Homol-D in some lineges. This scenarior is possible mainly through the intermediate stage where both Rap1 and Homol-D were used.

(b) The possible ways of regulatory change for a target gene:
   - Augmentation: add another TF
   - Abridgement: one of the original TFs no long target this gene
   - Switching: switch from one TF to another TF. Possibly involves the ”buffering” state where both TF target the same gene

(c) Question. coordinated gain of Rap1 BS in 35 genes? And in particular, in close proximity to the existing Homol-D BS in these genes?

7. Evolution of motif using in yeast TRN [Ihmels & Barkai, Science, 2005]

   Problem: change of GRN underlying the change of expression pattern of aerobic genes.

   Background:

   (a) The main function of mitochondia is the production of ATP through aerobic respiration, oxidation of glucose, pyruvate, and NADH. This process is dependent on the presence of oxygen. When oxygen is limited, the glycolytic products will be metabolized by anaerobic respiration, a process that is independent of the mitochondria.

   (b) Whereas growth of most yeast species requires oxygen, Saccharomyces cerevisiae (SC) grows rapidly in its absence and prefers to ferment glucose anaerobically even when oxygen is present.

   (c) In SC, the stress-related genes (STR) are induced during the relatively slow growth in nonfermentable carbon sources, such as glycerol or ethanol, which requires mitochondrial function.

   Results:

   (a) Gene expression patterns in SC and CA: RP (ribosomal proteins), rRNA genes, are activated during exponential growth phase in both SC and CA. MRP (mitochondria ribosomal proteins) are also activated with RP and rRNA in CA, but not in SC. In fact, MRPs are activated with STR in SC.

   (b) Change of cis-regulatory sequences of MRP: candidate motif *AATTTT* (RGE, rapid growth element). RGE is present in rRNA, RP in both SC and CA, and MRP in CA, but no in MRP in SC (hence named growth element).

   (c) Evolutionary history: RGE is presented in all species before genome duplication event, suggesting that the changes of RGE in MRP genes were due to loss events after genome duplication.

   Discussion: support the hypothesis: motif loss → change of expression of MRP genes → phenotypic change

8. Differential Cluster Algorithm (DCA) [Ihmels & Barkai, PLoS Genetics, 2005]

   Motivation: phenotypic difference between 2 yeasts, find the genetic basis in terms of expression

   Background: the difference between Calb and Scer resulting from different environments

   (a) Morphology and cell division: Calb may grow as hyphal cells (invasive form) by a non-budding mechanism. Note Spom is a fission yeast with non-budding mechanism.

(b) Respiration: fast growing Scer cells use fermentation, while Calb cells rely on respiration. In addition, fast respiration may lead to oxidative stress.

(c) Host nutrients: Calb may obtain different, specific subsets of AAs from the hosts.

(d) Host defense system: Calb needs to protect again host immune cells (e.g. macrophage).

Methods:

(a) Data: Calb data of 244 expression profiles, including stress response, mating type, growing in hyphal cells, etc.

(b) Differential clustering analysis: cluster genes in one species $\Rightarrow \forall$ cluster; check its pattern in the second species: whether it is split into 2 clusters, same cluster, etc. Thus a cluster in one species falls into 4 groups: conserved, partially-conserved, split, not conserved.

(c) Higher-level analysis of co-regulation patterns: group modules (co-expression) into module-tree, and see the patterns of association among modules.

(d) GO-connectivity network: connect two GO if their members belong to one transcriptional module.

Results:

(a) Overall pattern of conservation/divergence of co-expression: fully conserved clusters: ribosomal gene, oxidative phosphorylation; partial or split: protein synthesis, hexose metabolism, GCN4 targets; not conserved: cell cycle, RNA splicing, transcriptional regulation.

(b) Cell cycle: highly divergent. The genes involved in S-phase to mitosis transition are co-expressed in Calb, but not in Scer. The Scer CDC28 (major CDK) and CLB2 (cyclin) clusters are missing in Calb. Probably reflecting the difference in morphology and cell division.

(c) Amino acid biosynthesis: highly coexpressed in Scer, but split into four clusters in Calb: Arg cluster, aromatic AA cluster, Met cluster and general AA cluster. Also, additional motifs in the 4 clusters in Calb (other than Gcn4), in particular AATTTT motif in three clusters except Met cluster (the motif is also enriched in ribosome biogenesis genes [Ihmels & Barkai, Science, 2005]). Probably reflecting different AA availability. Also Arg cluster is involved in defense again host macrophages.

(d) Module tree and GO comparison: (1) AA biosynthesis and protein biosynthesis are associated in Calb, but not in Scer. (2) Mitochondrial ribosomal module is associated with protein biosynthesis in Calb, but not in Scer. Probably reflecting that Scer fermentation does not depend on Mitochondrial. (3) Carbohydrate metabolism and stress response are associated in Calb but not in Scer. Probably reflecting oxidative stress due to respiration in Calb.

9. Inter-species divergence of gene expression in yeast [Tirosh & Barkai, NG, 2006]

   Problem: what are the factors that determine the expression divergence?

   Hypothesis: the gene expression of TATA-box containing genes evolves faster.

   Methods:

   (a) Data: 4 yeast species (1 species has 2 strains); cDNA microarray using Scer coding sequences; 32 conditions under environmental stress (2 stress + 5 stress conditions * 6 time points/condition)

   (b) Defining expression divergence (ED) of a gene: the biases include (i) different responses/kinetics of different species; (ii) scaling difference of different arrays; (iii) large errors of with genes with large responses. The ideas for dealing with these biases:

      - Data normalization: the log2 expression will be normalized, using the the mean and standard deviation of the corresponding array.

- Pairwise ED: based on the expression profiles (32 conditions) of the orthologous genes in two species. Similar to Euclidean distance, but add an normalization factor that corrects for the large errors of genes with high expression ratio. The parameter of this correction term (controling the need of correction) is determined by testing the ED of orthologous (should be high) vs non-orthologous (should be low) genes.
- Data/time point selection: when comparing two species, only select the time points where the two species are highly correlated.
- Correction of pairwise ED: subtract ED with the average intra-species divergence (divergence between different strains) of the two species.
- Overall ED: each pairwise ED will be normalized (using the mean and standard deviation of all genes of this pair of species), then average of all pairwise ED.
- ED normalization: ED is not normally distribution, thus define $ED' = \log(ED + k)$.

Results:

(a) Transcriptome variation and divergence: (i) intra-species: r = 0.924 for individuals of the same species; (ii) inter-species: r = 0.679 for comparison of different species.

(b) Transcriptional plasticity (the degree of change across conditions), intra-species variation and inter-species divergence are significantly correlated. This suggests some general genetic mechanism underlying expression evolution.

(c) TATA-containing genes have higher expression divergence than non-TATA genes:

- Functional groups with more TATA genes have higher expression divergence: 1) stress-related genes, membrane proteins, AA metabolism genes tend to have more TATA and have higher ED; 2) cytosolic ribosomal proteins, genes involved in translation regulation tend to have fewer TATA and lower ED.
- Within a single functional group: TATA genes have higher expression divergence, e.g. stress-related genes.

(d) TATA-containing genes show higher ED across a number of datasets and organisms (mammals, drosophila, worm).

Discussion:

(a) TATA-promoter is more conserved than non-TATA promoter. Hyp: TATA genes have more TFBS, and this leads to higher conservation.

(b) Why TATA genes have higher ED? Hyp: TATA containing genes may be associated with different transcriptional complexes; they tend to amplify noise, thus tend to occur in genes whose expression noises are easily tolerated.

10. Identifying cycling genes [Lu & Bar Joseph, Bioinfo, 2006]

Problem: find cycling genes using expression data (periodic expression pattern)

Idea: true cycling genes tend to have periodic expression in multiple species (conservation of periodic expression pattern)

Background:

(a) comparison between budding yeast and fission yeast: a small set of core genes that are periodically expressed in both organisms. However, this observation may suffer from many artificts: difference of the sources of microarray data; difference of the methods to detect periodicity, etc. Methods:

(b) each homologous gene has a periodicity score in one specis. The gene with high scores in all species should be ranked higher. Modeling the uncertainty of sequence homology, allowing one gene to have multiple putative homologs.

Remark: use conservation of expression pattern - not the expression levels, but the higher-order characteristics of expression (periodicity, etc.)

11. Rewiring of yeast mating type genes [Tsong & Johnson, Nature, 2006]

    Backgroud: yeast has two mating types, a and α. Generally, expression of a-specific genes (asgs) leads to type a; and expression of α-specific genes (αsgs) leads to type α. But the mechanisms are different in different species:

    - C. albicans (Ca): (likely ancestral mechansim) MAT $a2$ expression, acticates asgs, in a cells; MAT $\alpha 1$ expression, activates αsgs, in α cells.
    - S. cerevisiae (Sc): asgs are constitutively active; but in α cells, they are repressed by $\alpha 2$, and αsgs are activated by $\alpha 1$.

    Problem: regulation of asgs has changed during evolution, despite that their expression patterns (only in a cells, but not in α cells) not. Specifically, the major transitions are:

    - asg expression becomes independent of $a2$;
    - asg comes under negative control of $\alpha 2$.

    How do these changes occur while maintaining expression patterns? At cis- or trans- level? How the fitness barriers are overcome?

    Results:

    (a) Identification of asgs in Ca: six genes, a-cell specific expression and orthologs of Sc asgs.

    (b) Evolution of cis-regulatory elements of asgs: (i) in C. albicans and related species: $a2$ binding site and Mcm1 binding site, separated by 4 bp; (ii) in S. cerevisiae: $\alpha 2$ binding site, Mcm1 binding site, and another $\alpha 2$ binding site; (iii) at intermediate species (K. lectis), both $a2$ and $\alpha 2$ sites are found in CREs. Thus the cis-level change is: loss of $a2$ sites, with gain of $\alpha 2$ sites (or simply conversion from $a2$ to $\alpha 2$ sites since they are very similar).

    (c) Emergence of $\alpha 2$-Mcm1 interaction: the Mcm1-interacting interface of $\alpha 2$ is not conserved around Ca, but conserved in Sc and neighbors. The evolution of $\alpha 2$-Mcm1 interaction (the AA sequence of $\alpha 2$) occurs in K. lactis lineage, coincoiding with the change in the cis-regulatory element of K. lactis.

    Discussion: the expression patterns of asgs were conserved during evolution despite changes. This is achieved by (Figure 6):

    - Tuning up Mcm1 activitity, thus making a2 unnecessary for activation in a cells;
    - Meanwhile, the evolution of Mcm1-$\alpha 2$ interaction allows expression to be repressed in α cells.

    Together, this allows change from positive control by a2 to negative control by $\alpha 2$. As evidence, in the intermediate (Kl lineage), both positive and negative controls are present (redundant state): a2-Mcm1 interaction allows asgs to be expressed in a cells; meanwhile, $\alpha 2$-Mcm1 interaction represses expression of asgs in α cells. The changes are expressed as:

    - "Tuning up" of a binding site for Mcm1 (because of the increase of AT content in the surrounding region), making gene expression independent of $a2$;
    - A small change in existing $a2$ binding sites, converting their recognition from $a2$ to that of $\alpha 2$;
    - A small change in the amino-acid sequence of $\alpha 2$, allowing it to bind DNA cooperatively with Mcm1.

Question: in the CREs of asgs after the changes, the first $\alpha 2$ site is similar to the earlier $a2$ site (thus could be created by a few substitutions), but there are other $\alpha 2$ sites in the other side of Mcm1 site, which will need many more mutations. How could this happen?

Remark:

(a) To understand evolution of a system, first need to understand the functional constraint or new demand (adaptation) of the system. What is the objective that evolution is optimizing?

(b) How do large changes happen while maintaining the functional need? An initial neutral or weakly deleterious change could: (i) makes another part of the system unnecessary (neutral); or (ii) be compensated by other changes in the system (weakly deleterious change). A chain of such changes could happen: $\Delta x \to \Delta y \to \Delta z \cdots$, which could lead to eventually large changes of the system.

12. TFBS gain and loss in Yeast [Doniger & Fay, PLoSCB, 2007]

Methods:

(a) Data: 4 yeast species, 3,761 integenic sequences after filting regions with high indels and high missing data.

(b) Classification of conserved and semi-conserved sites: conserved (HB model in all lineages), semi-conserved (HB model in all but one lineage, neutral in that lineage). Note: semi-conserved model only detects TFBS loss

Results:

(a) Of all constrained TFBS: 2/3 are conserved, and 1/3 are semi-conserved (estimated using a mixture model of conserved and semi-conserved)

(b) Lineage specific loss and turnover: only 40-60% loss can be explained by compensatory turnover.

(c) Effects of semi-conserved BS on gene expression: 8/11 semi-conserved sites (no turnover) have no effect on gene expression

(d) BS gain: more than 50% experimentally identified BS are not conserved or semi-conserved

13. TFBS divergence in Yeast [Borneman & Snyder, Science, 2007]

Problem: are TF-binding events conserved across species? Can the conservation/difference be explained by the sequences?

Methods:

(a) Data: 3 closely related yeast species, Scer, Smik and Sbay (s.t. there is no alignment ambiguity); 2 TFs (Ste12 and Tec1) known to be involved in pseudohyphal conditions (nitrogen limitation)

(b) Identify TFBS from ChIP-chip: on 3 species. A single binding event is defined as a discrete peak (that exceeds the threshold). The size of this event may not be small, e.g. 1-2kb (from Liu & Brutlag, NBT, 2002; and from visual inspection of Fig. 2)

(c) Predict TFBS from PWM: PWM is constructed using MDScan

Results:

(a) TFBS diverge substantially in three yeast species. Test: (i) compare multiple classes of TFBS: those conserved in all 3 species; only 2; and only 1. Only about 20% binding targets were conserved in all 3 species. (ii) quantatitive difference of the strength of binding.

(b) Sequences of conserved binding regions: (i) Tec1: for all sites with conserved bindings, 83% of them contain PWM matches; and 67% of 2/3 conserved sites contain PWM matches and 71% for 1/3 conserved sites. (ii) for Ste12: very low portion of experimentally bound sites contain PWM matches. Only 24% of regions with conserved binding in all three species contain a significant Ste12 motif. It is likely that Tec1 brings Ste12 to its targets.

(c) Sequences of partially conserved or lineage specific binding: (i) In about 14% (Tec1) and 10% (Ste12) binding events, the lack of conservation of binding is accompanied by the lack of sequence conservation. (ii) In 45 (Tec1, 12%) and 9 (Ste12, 3%) instances where a PWM match occurred in all three species but where that region was experimentally bound in only two species.

(d) Binding site conservation and function: among the promoters that are likely functions (genes show altered expression in the relevant condition), those with conserved binding or conserved motif matches are not enriched.

(e) Some divergence of TFBS can be explained by the change of functions of the target genes.

- Target genes are often TF, and the conservation of binding for these genes are very high.
- Ste12 binding loss in Scer (conserved binding in Smik and Sbay but not in Sc): enriched with mating genes. These genes still contain Ste12 binding sites, and presumably functional under mating conditions. Thus: Ste12 binds to these genes under filamentous growth condition in the ancestor, but binds under mating conditions in Scer. Examples: Chs1 (chintin biosynthesis), Fus3 (inhibit filamentous growth under mating).
- Ste12 binding gain in Scer: enriched with carboxylic acid transport genes.

14. Evolution of galactose-metabolism regulation in yeasts [Martchenko & Whiteway, Curr Biol, 2007]

Background:

(a) Galactose catabolism in S. cerevisiae: galactose is converted to glucose-6-phosphate through the Leloir pathway, including Gal1, Gal10, Gal7 (operon) and Gal5. The pathway is activated through: galactose induces Gal3, which bind with the repressor Gal80, and releases it from the transcriptional activator Gal4. In addition, the transcriptional repressor Mig1 stops the transcription of these genes in the presence of glucose.

Results:

(a) Conservation of galactose regulatory genes: Gal3 is missing in Calb, Gal80 has low sequence similarity (40%), Gal4 has high similarity in DNA-binding domain, but Calb Gal4 encodes a much smaller protein (261 AA vs 881 AA of Scer) with Gal80-interacting domain missing.

(b) Promoter sequences of Gal enzymes: (Figure 1) Scer: Mig1 and Gal4; Klac: Mig1, Gal4 and Cph1; Calb: Cph1. Cph1 is an ortholog of Scer Ste12. Also note that in the Gal1/10/7 promoters, four copies of palindromic motifs (half of the motif is similar to Ste12), not present anywhere else in the Calb genome.

(c) Galactose catabolism genes Gal1/10/7 are induced by galactose in Calb (thus conserved expression pattern).

(d) The role of Cph1 in Gal regulation: use Gal10 promoter activity in response to galactose to analyze the effects of regulation. Cph1 is required for Gal10 regulation: in Cph1 mutant, Gal10 activation is much lower with galactose induction, and this effect depends on Cph1 binding site.

(e) The different function of Gal4 in Calb: identify targets in Gal4 mutant strain through gene expression. The affected genes include glycolytic genes and subelometric (TLO) genes.

Discussion: the difference in the regulation of the Leloir-pathway genes may be due to the fact that galactose plays important roles in C. albicans adhesion and biofilm formation, processes that contribute to the virulence of this pathogen and which are absent in S. cerevisiae.

15. Adaptive conflict in galactose pathway [Hittinger & Carroll, Nature, 2007]

Background:

- In Scer: galactose induces the galactose utilization pathway (Gal1 is the first enzyme), through Gal3 (co-inducer), Gal80 (repressor) and Gal4 (activator).

- In Klac: similar switch, however, only KlacGal1, instead of Gal1 and Gal3.

Model:

- Adaptive conflict in ancestor: only one gene Gal1/3, moderate induction by galactose. The conflict exists because: Gal1 needs to be highly inducible by galactose, while Gal3, being the sensor of galactose, need not have a high induction by galactose.
- Resolving conflict by gene duplication: Gal1 and Gal3, each evolving specialized function; in addition, the change of promoters (Gal4 binding site configuration) allows Gal1 to be strongly induced.

16. Promoter and expression divergence [Tirosh & Barkai, MSB, 2008]

    Motivation: can expression divergence be explained by change of TFBSs?

    Methods:

    (a) Data: (1) human and mouse tissue-specific expression; (2) human-chimp liver expression; (3) yeast stress-responses; (4) yeast mating responses in Scer, Spar and Smik - only differential expression (up or down).
    (b) Promoter sequence divergence: if a motif is present in some species, but not the other ones.

    Results:

    (a) In general, changes in TF-binding sequences are not correlated with expression divergence (ED): in datasets (1)-(3), however, many limitations: (1) more distant enhancers in mammals; (2) combinatorial regulation; (3) TF binding sequences may not indicate binding.
    (b) More controlled setting: yeast mating response, where Ste12 is the main regulator. Ste12 BS divergence is correlated with ED. In addition, seven other TFs may contribute (Tec1, Swi6, Mbp1, etc.). Total, about half of the differential up-regulation of genes with Ste12 sites can be explained by the TFBSs.
    (c) Additional unexplained differential upregulation: the divergence of flanking sequences of Ste12 TFBSs (40bps in each side) are correlated with additional ED; in three cases, this means differential nucleosome occupancy (predicted by [Segal06] model).

17. TF substitution during the evolution of yeast ribosomal gene regulation [Hogues & Whiteway, Mol Cell, 2008]

    Problem: RP (ribosomal proteins) are regulated by Rap1 in S. cerevisiae (Sc), but this regulation is specific to Sc [Tanay, 2005]. What is the mechanism of regulation in other yeast species?

    Summary: Tbf1 and Cbf1 are regulators of RP genes in C. albicans (Ca) and ancestor, thus during evolution, the regulator has been switched to Rap1 in the Sc lineage.

    Results:

    (a) Tbf1 and Cbf1 motifs are enriched in RP orthologs in Ca, while Rap1 not.
    (b) Tbf1 and Cbf1 bind to RP protomters in ChIP-chip experimens. The presence of Cbf1-binding sites alongside only half of Tbf1 sites incidates that Cbf1 is not required for RP specialization of Tbf1 (probably facilitate access of Tbf1).
    (c) Tfb1 is an essential activator of RP expression: Tfb1 mutant shows dramatically reduced expression of RP.

    Discussion:

(a) Tbf1 and Rap1: both are Myb domain proteins, universally found at euk. telomeres, and they also physically interact. In all yeast species, Rap1 replaces Tbf1 function in telomere; and then before speciation of K. lactis, Rap1 substitutes Tbf1 at RP promoters. There may be selective pressure to connect the two functions: telomere stability and RP expression.

(b) Model of [Tanay05]: the ancestral species has Homol-D and IFHL motifs; in S. cerevisiae and its relatives, HomolD is replaced by Rap1, and IFHL accumulates some mutations and becomes Homol-E. The new model: the ancestral species has Tbf1 and Cbf; in S. cerevisiae amd relatives, Tbf1 (Homol-E) is replaced by Rap1, and accordingly, Cbf is replaced by IFHL (both Cbf1 in C. albicans and IFHL in S. cerevisiae seem to play supporting role to Tbf1 and Rap1, respectively).

Remark: the function of Homol-D box in other yeast species (not found in C. albicans) is not discussed. Perhaps related to Tbf1-Cbf function in the ancestral species, but as Tbf1 is replaced by Rap1, Homol-D box is no longer needed.

18. Evolution of Mcm1 regulation in yeasts [Tuch & Johnson, PLoS Biol, 2008]

Problem: the evolution of Mcm1 controlled regulatory relations.

Summary: conservation of Mcm1 targets; the change of Mcm1 targets caused by: 1) change of Mcm1 binding sites; 2) change of Mcm1-cofactor interactions.

Methods: experimental determination of Mcm1 targets by ChIP-chip in S. cerevisiae (Sc), K. lectis (Kl) and C. albicans (Ca).

Results:

(a) Most Mcm1-cofactor interactions are conserved, but their individual target genes have changed dramatically.

(b) Ribosomal genes: new Mcm1 targets in Kl lineage (about 70 genes), and three other (unrelated) lineages, but not the other ones. Also, these new Mcm1 targets have companion changes of Rap1 binding sites.

(c) Mcm1 targets in Kl but Ca, but no in Sc: Arg biosynthesis genes. Lineage-specific loss in Sc, probably driven by gene duplation: Mcm1 to Arg80, and Arg80 may take over the role of Mcm1.

(d) Non-canonical Mcm1 motifs in genes involved in white/opaque switching in Ca lineage (specific to Ca, adaptive to host interaction): also accompanied by new Wor1 binding sites.

19. Chemical stress response of Scer and Cgla [Lelandais & Devaux, GB, 2008]

Background: Cgla is a pathogenetic yeast with some drug resistence.

Methods:

(a) Data: transcriptional response to benomyl (an antifungal agent inducing oxidative stress in Scer), measured in 2,4,10,20,40 and 80 min.

Results:

(a) The similarity/conservation of transcriptional profiles: (1) PCA analysis: the Cgla response is slightly faster; (2) conservation of motifs: AATTTT motif, STRE (stress response element) are all conserved in co-expressed genes.

(b) Different roles of Yap1: ScYap1 regulates about 40% of up-regulated genes, while CgYap1 only 25%, and the overlap is even smaller (14 genes out of about 200 up-regulated genes in both species).

(c) Different binding properties of Yap1 in Scer and Cgla: Scer favors TTA[C/G]TAA, while the dominate Cgla motif is TTACAAA. In addition, the promoters diverge very fast between two species. Promoter analysis of other yeast species suggest that: ancestral sequences are regulated by TTA[C/G]TAA, but the regulation is gradually reduced during evolution.

Discussion: an example of conservation of expression pattern with the divergence of the underlying regulatory network. The conservation may be achieved by: (1) co-evoluation of TF binding property and TFBS; (2) the regulatory network, e.g. other TFs play the role of Yap1 (Msn2/4 or other Yap1 paralogs).

20. Evolution of carbohydrate metabolism regulation in Calb [Askew & Whiteway, PLoS Pathogen, 2009]

    Background: response to glucose signaling in Scer, e.g. switching to glucose from a nonfermentable carbon source:

    - Activation of glycolysis enzymes: Gcr1/2 and Tye7 are activators (Tye7 plays a less important role). In addition, Rap1, Abf1, Reb1 also play roles.
    - Promotion of fermentation: induction of PDC expression.
    - Glucose repression: of catabolism of other sugars (e.g. galactose) and of respiration (pyruvate dehydrogenase or PDH, TCA cycle), via Mig1 and Rgt1.

    Background: regulation of glucose metabolism in Klac

    - Gcr1/2 has orthologs in Klac (but not in most other species).
    - Glucose repression circut (of TCA cycle) is missing in Klac.

    Model: regulation of carbohydrate metabolism in Calb (Figure 1):

    - No Gcr1/2 ortholog in Calb. The main regulators are Tye7 and Gal4.
    - Glycolysis: Tye7 increase expression of glycolysis enzymes with the assistence of Gal4. Tye7 helps to commit cells to glycolysis.
    - Fermentation vs respiration: controlled via two enzymes PDC (to fermentation) and PDH (to respiration) and TCA cycle. PDH is activated by Gal4 and TCA cycle is not repressed.

21. Gene expression divergence is coupled to nucleosome organization [Field & Segal, NG, 2009]

    Hypothesis: change of nucleosome organization of DNA sequences is one way of achieving different transcriptional patterns.

    Methods:

    (a) Gene sets: defined by GO categories, orthologs in both Scer and Calb.
    (b) Defining conservation of gene sets: if a set is correlated with CRP (cytosolic RP) genes in both species, then category I; anti-correlated in both, category II; correlated in Calb but not Scer, category III.

    Results:

    (a) Category I genes: often growth-related; II genes: stress, mating type, energy reserve, autophage, etc.; III: MRP, electron transport and TCA cycel.
    (b) Nucleosome organization of genes: (1) category I: relatively open in both Scer and Calb; (2) Category II: relatively closed in both Scer and Calb; (3) Category III: relatively open in Calb, but closed in Scer.

    Discussion: adaptation to anaerobic (fermentative) lifestyle in species after WGD eliminates most of the need of respiration-related genes, thus lower their basal expression level through changing nucleosome occupancy.

22. Evolution of oxygen responding system in yeast [Fang & Bao, PLoS ONE, 2009]

    Background: [Merico & Compagno, FEMS, 2009]

- Scer can grow well in the absence of O_2. The key regulators are Rox1, Hap1 and Mot3. Rox1 (heme-dependent) repression of hypoxia-genes is a key regulatory mechanism.
- Klac is aerobic species, grow poorly in the absence of O_2.
- The inability of growtin in anaerobic condition of Klac: lack of an efficient mechanism to maintain a high glycolytic flux (reduced flux in the pentose phosphate pathway, which is important for Klac), and to balance the redox homeostatsis under hypoxic conditions. KlGcr1 (glycolysis activator) is up-regulated by hypoxic condition, but only transient effect.

Results:

(a) Scer paralogs (from WGD): enriched with anaerobic-aerobic pairs, where one gene is expressed under aerobic condition and the other under anaerobic condition. Ex. Hyp2/Anb1 (heme biosynthesis), Cox5a/Cox5b (respiratory chain biogenesis).

(b) Klac often has one ortholog of these anaerobic-aerobic pairs, and the expression of Klac orthologs resemble the aerobic version of Scer.

(c) Rox1 protein sequence: low similarity to Scer Rox1, except for the DNA binding HMG domain, the structure features are quite different.

(d) Rox1 binding sites in the oxygen responding genes: mostly missing in Klac.

Discussion: Klac does not have the Rox1-mediated oxygen responding system. The evolution of this system in Scer: WGD, and one copy may acquire a Rox1 binding site, and become subject to the Rox1 control.

23. Expression evolution using yeast hybrid [Tirosh & Barkai, Science, 2009]

Problem: distinguish trans- and cis- effect on the evolution of gene expression.

Methods:

(a) Yeast hybrid analysis: suppose the two parents have two different alleles for a gene of interest: A and a. In general, if expression of A in one parent and a in another parent are different, we do not know whether the difference is due to trans- or cis- divergence between two parents. Create hybrid and measure allele-specific expression of A and a in the hybrid:

   - The difference between A and a in two parents: the combined effect of cis- and trans-.
   - If A and a expressions are different in hybrid: the diffence is only due to cis-, because the trans- environment is the same.
   - Subtracting the expression diffence in hybrid from the difference in parents: the difference must be due to trans-.

(b) Data: Scer and Spar microarry under 4 conditions: rich media, heat shock, Rpd3 inhibitor, trichostatin A (TSA).

Results:

(a) cis-effects account for most of expression divergence across all conditions; and trans-effects are condition specific.

(b) Trans-effects correlat with expression divergence (in other studies), i.e. genes with high trans-effects tend to have high expression divergence. These genes also tend to have TATA box and lack nucleosome-free region.

(c) Trans- effects correlate with the interpretation of environmental signals: i.e. genes with high trans-effects tend to show larger expression changes under different environmental conditions.

(d) About 20% hybrid specific expression (expression level in hybrid is higher or lower than in either parent) are from compensating changes, i.e. trans- and cis- effects are in the opposite direction, thus compensating each other. This suggests purifying selection of expression (thus parents are similar, but when in hybrid, different trans- environment could lead to different expression).

24. Phenotypic profile of Calb TFs [Homann & Johnson, PG, 2009]

Background: Calb colonies are complex structures consists of three types of cells: budding yeast, pseudohyphae and hyphae (the last two: invasive growth that penerate into the solid medium). Two extreme form of structure:

- "Wrinker" structure: consists of all three types. Extensive extracelluar matrix deposition.
- "Smooth" structure: primarily yeast cells, absence of an extensive extracelluar matrix.

Methods:

(a) Transcriptional regulator knockout library (TRKO): two independent strains for each TF (require two rounds of disruption, since Calb is diploid. The process may create unintended mutations, thus two strains). A total of 317 strains representing 143 TFs (consistent phenotypes on duplicate strains).

(b) Phenotyping: on 55 conditions.
   - Condition: different nutrients, temperature, stresses, antifungal druges.
   - Measurement: growth and morphological phenotypes (wrinking or invasion).

Results:

(a) TOR pathway: whether a TF is controlled by TOR pathway is tested via the phenotype of TF deletion on rapamycine and/or caffeine (TOR inhibitor). The core TOR regulatory network is highly conserved between Scer and Calb (most Scer TFs controlled by TOR show rapamycine/caffenine phenotype).

(b) Iron acquision and homeostasis regulation: the source and abundance of iron vary greatly with microenvironment and iron acquision and homeostasis is a special challenge for Calb. The main regulator in Scer is Aft1/2 (positive), while in Calb, it is Sfu1, a negative regulator.

(c) Colony morphogenesis and invasive growth: a number of new TFs are identified that control colony morphology. Ex. Gat2 regulats formation of colony wrinkling, ECM production and invasion; other regulators appear to link specific cues in environment to colony phenotype.

(d) Comparison of TF functions in Scer and Calb: many have conserved phenotypes, some exceptions. Gal4: galactose utilization in Scer but not in Calb; Rtg1: glutamate auxotrophies in Scer but not in Calb; Met31/32: methionine auxotropy (double deletion) in Scer, but no phenotype in Calb.

Remark:

- There are both positive and negative regulators of iron homeostasis in Calb, how does one know Sfu1 plays a similar role as Aft1/2 in Scer? There may also be other positive regulators in Calb (or unidentified negative regulators in Scer), then one cannot say it is a switch from positive to negative regulation in two speices.
- The difficulty of comparing phenotypes of orthologous TFs in different species: conditions may not be exactly the same; baseline sensitivities to environmental cues may be different; gene duplications and loss.

25. Evolution of RP regulation [Lavoie & Whiteway, PLoS Bio, 2010]

Problem: the function of Tbf1 and Cbf1 in regulating RP in Calb (same signaling pathways), and the role of other proteins.

Results:

(a) Protein conservation: Crf1 has no ortholog in Calb (a recent appearance in Scer lineage).

(b) Binding targets of main factors: mapped through whole-genome ChIP-chip. Hmo1, Rap1, and Tbf1 have no significant overlap between the two species. In particular, Rap1 binds far more targets in Scer than in Calb.

(c) Functions of main factors according to the GO categories of the bound targets: four generalist Hmo1, Rap1, Tbf1 and Cbf1; two specialists Fhl1 and Ifh1.

- Cbf1: centromere in Scer, but not in Calb.
- Hmo1: essentially absent from the RP regulon in Calb.
- Tbf1: bind to telomeric regions; and moderate enrichment of cell cycle progression.
- Rap1: in Scer, in addition to RP and telemere, also in glycolysis and silent mating type locus. In Calb, telemere only.

(d) Evolution of binding specificity: Rap1 and Tbf1 both bind to longer motifs in Calb.

(e) Fhl1 and Ifh1 binds almost exclusively to RP promoters in Calb (conserved). Binding depends on Tbf1, but the peak coordinates of C. albicans Fhl1 occur at 37+/-25bp from Tbf1 peaks; while Fhl1-Rap1 about 96bp in Scer.

(f) The signaling-dependent Ifh1 association, histone modifications, and probably the recruitment and dissociation of the histone acetylation/deacetylation machinery at RP promoters occur in both species despite the remodeling of the ribosomal TF complex.

Discussion:

(a) The cis-regulatory changes allow coupling/decoupling of different regulons: e.g. RP and glycoylic genes are coupled in Scer through Rap1 binding; RP, electron transport chain, and sulfur starvation regulons mediated by C. albicans Cbf1.

(b) How would change happens? The initial changes of a few RPs, then rapid changes at other RPs to maintain stoichiometry.

Remark: one hypothesis is RP expression changes may be neutral, but rewiring may be driven by selective pression on adapative changes of co-regulation.

26. Gene duplication in RP regulation [Wapinski & Regev, PNAS, 2010]

Results:

(a) Ifh1 and Crf1 are paralogs, arising after WGD. The sequence analysis suggests that the ancestor is more like Ifh1. Furthermore, Cgla loss Crf1 gene (also loss RP duplicates).

(b) RP expression in six yeast species (both pre- and post-WGD) under stress: highly conserved patttern, the only exception is Cgla, whose RP expression does not change upon stress. For Calb, RP expression does not change at 37 degree heat shock, but change significantly at 42 degree heat shock.

Discussion/Model:

- RP expression patterns (in response to nutrients and stress) are generally conserved, despite the Tbf1/Rap1 switch in Scer subgroup.
- After WGD, Ifh/Crf gene in the ancestor duplicates and each paralog specializes its function and becomes Ifh1 (activator) and Crf1 (repressor) respectively. This is probably driven be RP gene dosage: need of tigher control of RP genes.
- In Cgla (post-WGD): RP gene loss is accompanied by the loss of Crf1 gene (less need of tighter RP control), also this is human pathogen, probably different environment (e.g. more stable) may also contribute to the change of RP expression. Note that ribosomal biogenesis is still reduced by stress, thus RP expression may be regulated post-transcriptionally.

27. Balanced polymorphism at GAL pathway [Hittinger & Rokas, Nature, 2010]

    Question: mechanism of balanced polymorphism at multiple alleles? If the alleles are unlinked, then recombination would reshuffle the alleles.

    Results:

    (a) Japanase strains of Sklu: Gal$^-$ - GAL pseudogenes and cannot utilize galactose; Portuguese strains: Gal$^+$ - GAL genes and can utilize galactose. However note that Gal3 is missing even in Por strains.

    (b) One hypothese to explain the maintanence of two network states is the lack of genetic exchange between the two populations. If so, then the divergence of all other genes in the genome should be high (similar to GAL), but this is not true.

    (c) Explanation of balancing polymorphism:

        - Under galactose-rich condition (e.g. Portuguese): Gal$^+$ > Gal$^-$ in terms of fitness.
        - Under galactose-lacking condition (e.g. Japan): Gal$^+$ may have slight conditional fitness cost (for unknown reasons).
        - On average, the recombinations from the two states have lower fitness: e.g. Gal4$^+$ and Gal80$^-$ strain has constitutively active GAL expression.

    Remark: an example of epistasis, where the fitness of one allele depends on the genetic background (e.g. whether Gal4 is advantageous depends on if Gal3 and Gal80 are present).

28. Pheromone response in Calb [Sahni & Soll, PLoS Biol, 2010]

    Background:

    - White/opaque switching: only in Calb and closely related Cdub. The function of the system may be: formation of biofilm by white cells that facilitate mating between minority opaque cells.
    - Different pheromone responses in white and opaque cells: (1) opaque cells: mating, similar to mating response in a or α cells of Scer; (2) white cells: a smaller number of genes are induced, no mating, but biofilm formation. Correspondingly, the expression responses to pheromone of white and opaque cells are different.

    Model:

    (a) Pheromone response singaling in opaque cells: similar to Scer with Cph1 (Ste12 ortholog) being the main TF activating downstream genes.

    (b) Pheromone response singaling in white cells: the upstream signal pathway is identical to opaque cells, the TF is Tec1 instead of Cph1. In addition, Tec1 regulates biofilm formation genes, which contain Tec1 binding sites (the motif is slightly different from ScerTec1).

    (c) Evolutionary path: two critical steps: (1) Cek2 regulation of Tec1 in white cells; (2) Tec1 binding sites in biofilm genes. Note that (1) may already be present, as Tec1 is a target of Kss1 in the filamentous growth pathway in Scer, thus only small modification may be needed. Since filamentation genes (which are controlled by Tec1) can help biofilm formation, thus step (1) itself may be advantageous at the beginning.

    Remark: the main problem of this model is: the specificity of response, i.e. how to avoid Tec1 activation in opaque cells and Cph1 activation in white cells (since Cek1/2 now regulates both Cph1 and Tec1)? This may not be difficult as Scer can achieve similar specificity in pheromone vs. filamentous growth pathway.

3.4 Regulatory Evolution in Insects

Evolution of eve stripe 2 CRM (S2E: stripe 2 element) [Ludwig & Kreitman, Development, 1998]:

- Problem: the expression pattern of eve stripe 2 is conserved across Drosophila species? And what is the sequence basis (whether sequences are conserved or not)?
- Methods:
  - Expression pattern analysis: reporter assay in Dm of the S2E from each of the 4 Drosophila species
  - S2E alignment: extract S2E sequences from each species and do the alignment
- Results: the sequences have undergone significant despite the conservation of expression pattern
  - Expression pattern: the spatial and temporal patterns are conserved across 4 species, but the level of expression is different: highest in Dm.
  - Sequence analysis: Fig. 4
    - 3/16 sites are completely conserved
    - kr (6): 2 have substantial substitutions and 1 has small indels
    - bcd (5): bcd-3 is not conserved in at least 2 species (bcd-3 has been shown earlier to be important for the expression pattern in Dm)
    - hb (3): hb-1 is not conserved in ¿= 3 species, and hb-3 has small indels
    - gt (3): overall conserved, but all 3 have indels (gt-3 has a large indel)
    - Spacing: significantly changed in 4 species
- Analysis/Hypothesis:
  - The expression pattern is under negative selection and the expression pattern of TFs are also conserved (some experimental evidence)
  - The sequence changes are caused by fixation of slightly deletious mutations and adaptive mutations
  - The changes are made possible via the "robustness" of the enhancer function (due to TFBS redundency) and also the compensatory changes

Evolution of eve stripe 2 CRM [Ludwig & Kreitman, Nature, 2000]

- Methods:
  - Chimeric S2E construct from S2E(m) and S2E(p) ==¿ S2E(p1-m2) and S2E(m1-p2) (m: Dm; p: Dp)
  - Reporter assay of chimeric enhancers
- Results: posterior shift or expansion of the stripe expression pattern
  - S2E(m1-p2): posterior shift, may be caused by the absence of a Kr site
  - S2E(p1-m2): subtle expansion of both the anterior and posterior borders
- Analysis: compensatory changes within CRM s.t. the expression pattern is conserved. Stabilizing selection of CRM. For a quantative trait controlled by many loci that is under stabilizing selection, the average selection coefficient / locus is small and the rate of substitution can be quite high.

Evolution of hairy enhancer in fly [Kim, J of Exp Zoo, 2001]

- Problem: conservation pattern of CRM, in particular, whether the length of the inter-TFBS region is under stabilizing selection

- Methods:
  - Hairy enhancer of 7 Drosophila species, including both close and distance ones (ex. Dpse)
  - Testing of length variation: test coevolution of length of two/multiple regions

    $$F = \frac{Var(S_1 + S_2)}{(Var(S_1) + Var(S_2))} \tag{3.1}$$

    H0: S_1 and S_2 are independent. Under H0, F has a F-distribution
- Results:
  - Pattern: highly conserved blocks and extremley variable inter-block regions: 17 conserved blocks + 16 regions
  - Testing of molecular clock and comparison with species tree: (i) conserved blocks: non-clock; (ii) variable regions: clock, and the tree constructed using the length is consistent with the species tree; (iii) overall: non-clock (due to conserved blocks)
  - Length variation of inter-block regions. Hyp: if the total length is under (-) selection, then should expect negative correlation between the regions (compensatory effect)
    - No significant coevolution between adjacent inter-block regions
    - Some significant coevlution between some non-adjacent regions (at 3 end), but positive instead of neagive
  - Changes within the conserved blocks: show lineage-specific changes, in particular in Dpse lineage. Evidence of change of expression of hairy in Dpse, which might be due to the change of CRM

  Discussion/Conclusion:

  - Inter-block regions under no (-) selection
  - Lineage-specific changes on the conserved blocks are important

Conservation of TFBS in Drosophila species [Emberly & Siggia, BMC Bioinformatics, 2003]

- Methods:
  - Data set: 30 CRMs, 315 verified TFBS of early development
  - Alignment: LAGAN and SMASH. Parameters are varied to maximize the significance of the BS intersection with conserved blocks
  - Assessing TFBS conservation: binary measure: number of TFBS within conserved blocks; Intersection measure: bps of TFBS in conserved blocks
- Results:
  - TFBS are not significantly more conserved than (random) noncoding DNA (Figure 1). E.g. if use >= 60% identity as conservation cutoff → 50% sensitivity and 61% specificity (39% noncoding DNA will meet this threshold).
  - Module scanning for two Dm and Dp species and take the intersection → not significantly enriched for known modules

TFBS gain and loss in Drosophila early development genes [Dermitzakis & Clark, MBE, 2003]

- Data: CRM of 5 genes involved in AP patterning. Five fly species, Dmel, Dsim, Dyak, Dsech, Dore. Known binding sites of Bcd (57), Cad, Hb (97), etc.

- Polymorphism & substitution pattern: (i) reject neutral hyp. using Tajima's D: probably under (-) selection; (ii) clustering of polymorphic sites: fraction of regulatory sequences that is functional may be higher (not a small portion)
- TFBS conservation: Bcd and Hb binding sites. (i) most known Bcd and Hb BS are conserved (predicted to be functional) in all species, only Bcd-3 and Hb-1 of eve 2 are exceptions; (ii) many predicted BS, but many of them are not conserved (gained or lost), in Fig. 5

TFBS gain and loss in Drosophila early development genes [Costas & Vieira, Gene, 2003]

- Methods:
  - 7 transcription factors in D. melangoster: Bcd, Cad, Ftz, Hb, Kni, Kr and Tll.
  - Binding sites: experimentally verified (119) in 8 CRMs. Remove 15 sites whose PWM scores are below a cut-off.
  - 3 species: Dmel, Dpse and Dvir.
- Results:
  - The percentage of BSs present in Dmel, but not in another species in pairwise comparison: 23% with Dpse (time = 25 my); 37% with Dvir (time = 40 my).
  - PWM scores are similar for three groups (conserved in all three, in two, and only in melangoster) for most TFs. Exception is Hb. May be explained by: the change of TF; necessary to have a low affinity in one species; differece in background distribution

Abdominal pigmentation pattern in fly [Gompel & Carroll, Nature, 2003]

- Problem: genetic basis of abdominal pigmentation pattern evolution in fly
- Methods:
  - phenotype: abodminal pigmentation of 13 drosophila species. The pattern can be divided into 4 parts (see below)
  - expression of Bab2. Divide into 4 modes: midline elevation, midline repression, posterior repression and dimorphic (different patterns in males and females)
- Results:
  - Bab2 expression correlates with the phenotype in 9 out of 13 Drosophila species: for all 4 modes
  - When they are not correlated, Bab2 expression is correlated with trichome pattern (a related trait affected by Bab2) Discussion/Conclusion:
  - flexible change of TF expression is a general mechanism of phenotypic change (both convergence and divergence)
  - if a gene is involved in multiple traits, then they may conflict, i.e. the expression pattern is consistent with one trait, but inconsistent with the other trait (perhaps through other loci)
- Remark:
  - use binary vector of gene expression: to correlate with morphological trait. To test differential expression ⇒ convert contiuous expression to presence or absence
  - a phenotype may be expressed as multiple traits/characters/features: e.g. spatial pattern, temporal pattern. Correlate with the expression pattern which could also be spatial or temporal

Evolution of gene expression in the Drosophila melanogaster subgroup [Rifkin & White, NG, 2003]

- Problem: evolutionary change of expression patterns of genes involved in metamorphosis: extent of variation, which genes are more constrained, etc.?
- Methods:
  - data/experiments: metamorphosis (2 time points) of 4 Dmel strains and Dsim + Dyak; cDNA microarray
  - test of evolutionary mode: stable (negative selection), lineage-specific selection and neutral
- Results:
  - variation of gene expression in different lineages corresponds to phylogenetic pattern, i.e. closer lineages have smaller variation. Test: count number of genes whose expressions are significantly different in 2 lineages
  - majority of genes that have changed in at least one lineage are evolutionarily stable. Test: classify the genes by the evolutionary mode
  - comparison of different types of genes: (i) TFs are more often constrained than downstream targets (early developmental genes more conserved than terminal differentiation genes); (ii) genes that are induced in metamorphsis are more often constrained

Drosophila sex-dependent selection [Ranz & Hartl, Science, 2003]

- Problem/Hypothesis: sex-dependent selection is important for the difference of gene expression across species
- Methods:
  - data: male and female adult flies of Dmel and Dsim; cDNA microarray of Dmel
  - testing differential expression: a gene is differentially expressed in the 2 arrays (either between sexes or between species) if the confidence intervals do not overlap
- Results:
  - Prediction: sex dependent selection will lead to sex-specific evolutionary pattern (if a gene is sex-biased, and it changes its expression in evolution in one sex but not the other, then it will display a sex-specific pattern). Observation: genes that are differentially expressed in 2 species often have sex-specific evol. pattern. 50% genes are differentially expressed in 2 species, among which 83% show sex-specific pattern.
  - male-biased genes (high expression in male) have larger divergence than female-biased or non-sex biased genes. Test: divergence is defined by coefficient of variation across species
  - functional categories of genes showing differential expression between species: (i) no significant category in non sex-biased genes; (ii) in sex-biased genes: mating behavior; and phototransduction cascade in male-biased genes of Dsim.
- Discussion: reject the neutral model of expression evolution, which cannot account for the sex-specific pattern. Instead, sex-dependent selection drives the inter-species changes of gene expression

Evolutionary changes in cis- and trans- regulation [Wittkopp & Clark, Nature, 2004]

- Problem: expression difference in two species: cis- or trans- changes?
- Methods:
  - data: 29 genes DE in Dmel and Dsim

  - distinguish cis- and trans- changes: let Mel and Sim be expression of a gene in Dmel and Dsim respectively, and Mel(F1) and Sim(F1) be those of the species-specific genes in F1 hybrid. Then: (i) if cis-regulatory change, then Mel(F1) != Sim(F1) because trans-factors are identical; (ii) if trans-regulatory change, then Mel(F1)/Sim(F1) != Mel/Sim.
  - expression measurement: pyrosequencing from mRNA

- Results:
  - 28/29 (97%) genes show cis-regulatory changes
  - 16/29 (57%) genes show trans-regulatory changes. Thus overall: 12 genes only cis-regulatory changes; and 16 both cis- and trans- regulatory changes.

Wing pigmentation pattern in Drosophila [Gompel & Carroll, Nature, 2005]

- Problem: genetic basis of wing pigmentation pattern in Dmel vs Dbia (Dpse as outgroup): Dbia has a black spot in the anterior-distal region of the male-wing.
- Results:
  - Wing spot pattern is determined by the expression pattern of the yellow (y) gene
  - cis-regulatory change causes the change of spot pattern: GFP reporter expression under y upstream sequence (8kb) agrees with the spot
  - Wing element contains the cis-regulatory change (wing element: a known CRE of y): conservation of wing element in the 3 species; GFP reporter expression under the wing element recreates the spot
  - A segment of the wing element, called spot element contains the cis-regulatory change (675bp). Exp: wing → left and right parts. Only the left part reproduces the pattern, called spot element
  - Activation sequence lies in 335-530bp of spot element; and the expression in the posterior part is limited by the BS of Engrailed (2 BS)

CRE/CRM conservation in Drosophila: Dmel vs. Dpse [Richards & Gibbs, GR, 2005]

- Background:
  - Genome size: Dpse genoms is 17% larger than Dmel
  - Gene content: Dmel and Dpse shares about 10k genes (BBH); 12k Deml genes have ortholgous in Dpse; 1485 Dmel-unique genes, majority may be due to sequence gap, or low similarity
  - Alignment: 48% nts can be aligned. The average aa sequence identity of proteins is 77% with a mode about 85%
- Methods:
  - Data: 142 knowns sites of 30 genes, 63% upstream, 25% intronic, 6% downstream. The median length of binding site is 14bp and the modal position is 2kb from the putatitve TSS.
  - Procedure: compare the sequence identity among three groups: CRE, random intergenic control (RIC) and nearby
- Results:
  - CRE has a peak at $> 80\%$ and the mean identidy is 47.8% (= 72% identity in aligned CRE)
  - Nearby mean identity is 46.3% (= 66% identity in aligned CRE)
  - RIC mean = 42.4%

    - Difference is statistically signficant, but may be hard to detect: in average only 1 more position is conserved in CRE (14bp) than in nearby.

Enrichment of conserved blocks in CRM [Papatsenko & Dubchak, Genomics, 2006]

- Problem/Hypothesis: the distributions of conserved blocks in different types of sequences are different. This difference can be used for detecting functional sequences.
- Methods:
    - Data: 6 Drosophila species, including both close and distant ones. Types of sequences include: exons, regulatory sequences: promoters (close to TSS) and CRM, introns, UTR, and unannotated sequences
    - Pairwise vs multiple alignment: choose pairwise alignment because the blocks in multiple alignment are harder to interpret (e.g. perfected conserved in some lineages but not in other lineages)
    - Defining blocks: ungapped, 100% conserved
- Results:
    - Block distribution (frequencies of blocks of various sizes) in CRM is different from other types of sequences: (i) significantly different from all other types by a chi-square test; (ii) contain a larger amount of 20-30bp blocks, and also enriched in ¿20 blocks: in Dmel vs Dpsu, 40% of CRM contain 100% conserved blocks larger than 35-40 bp.
    - Block distribution in promoter regions: no preference for long ungapped blocks
- Discussion: future work include
    - Relax the definition of conserved blocks: allow small number of mismatches
    - Statistical analysis of the block-gap patterns: Markov chain model
    - Multiple alignment interpretation

Turnover of Zeste binding sites in Drosophila [Moses & Eisen, PLoSCB, 2006]

- Methods:
    - PWM of Zeste: constructed from 26 verified binding sites
    - Chip-chip experiment: 294 regions that bind Zeste (each region: 300bp around one peak); and align the 4 Drosophila species (within 10myr)
    - Predicting binding sites in D. mel: using PWM match cutoff ($p < 0.001$)
    - Classification of a predicted site (from PWM match) as conserved or not: LRT of HB model vs background (HKY) model
    - BS gain and loss rates: for any nonconserved sites (determined by the LRT based on HB model), assign its lineage of gain or loss. (i) loss rate: fraction of single lineage losses over the total number of ancestral sites (conserved + single-lineage losses); (ii) gain rate: number of single-lineage gain divided by the total sequence length.
    - Control sequences: 1,000-bp noncoding fragments located 23 kb on either side of the bound intervals.
- Results:
    - The effect of alignment error is neglibable via heuristics for correcting alignment error, checked via simulation with CisEvolver. Heuristic: search for BS match in one species, if the orthologous sequence is also a match and overlap by > 1 bp, then move the BS → often align nonorthologous sequences even for very close distance.

  - Bound regions contain a significant number of nonconserved functional sites: 1406 predicted sites in Dmel bound regions, out of which 215 D. mel sites are nonconserved. To account for FPs in these 1406 sites, compare with the number of sites in flanking region (count excess in bound region) → 806.7 functional binding sites, including 61.6 nonconserved functional sites. But the expected number of nonconserved sites in 807 functional sites is only 8.7, thus enrichement of nonconserved functional sites.
  - Binding sites gain and loss: predict BS in the bound region (1909) in all species, 1177 conserved.
    - Loss: 66 losess out of 1243 ancestral sites in bound vs 78 losses out of 841 ancestral sites in flanking regions. Reduced losses in bound regions.
    - Gain: 360 gains out of 223.499 kb in bound vs 602 gains out of 424.159 kb in flanking regions. Increase gains in bound regions.
  - Compensatory turnover model: 33 gains and losses (lineage-specific) occur in the same bound region. Comparing with the distribution of this co-occurrance among randomly permutated regions (?), it is not significant.

- Remark:
  - Study the binding site gain and loss from the perspective of natural selection: negative or positive selection of the number of binding sites. Selection is tested by comparing with flanking regions (negative control). Implication: compare the rates of different regions/groups of genes to detect selection
  - The common sources of error in the data: (i) nonfunctional binding sites (false positives): by treating all predicted sites as a mixture and estimate the number of true sites; (ii) alignment errors; (iii) TF specificity change in different species.
  - Application of LRT of conservation for BS gain and loss: if a BS is predicted to be conserved, then no need to check each orthologous site (thus an orthologous site is BS even if it may be weak) → increase the sensitivity of predicting BS

Large-scale analysis of CRM in Drosophila [Li & Halfon, GB, 2007]

- Problem: the sequence features & conservation pattern of CRM
- Methods:
  - Data source: 280 CRMs in REDFly (length $<$ 2.1kb): 19% in embryonic blastoderm, 13% in neuronal tissue. 61% in 5'; 13% overlap with promoter; 13% in 3'; 16% in introns.
  - Conseervation measure: pairwise alignment (DIALIGN), the fraction of aligned bases and PID
  - Test for homotypic clustering: (i) Fluffy-tail test: FTT score measures the overrepresentation of the strongest word; (ii) YMF: YMF score measures the most enriched word; (iii) supervised learning: classify CRM and non-CRM by their enrichment of common words
- Result:
  - GC content of CRM vs random non-coding DNA: GC content in CRM is higher than non-coding DNA (Wilcoxon test); a negative correlation between CRM length and GC content (Spearman's correlation)
  - Conservation of CRM vs random non-coding DN:
    - CRMs are significantly more conserved than random non-coding DNA
    - Sequences flanking CRMs are less conserved than CRM (significant, but the extent is small), but more conserved than random non-coding DNA

    - * The degree of CRM conservation decreases with increased evolutionary distance, but the difference between CRM and random non-coding remains constant
  - Homotypic clustering of CRM: only enriched in one subclass of CRM's (blastoderm).

Evolution of eve enhancers in sepsids and Drosophila [Hare & Eisen, PLoS Genetics, 2008]

- Methods:
  - Data: eve locus of sepsides, about 100 Myr away from Drosophila.
  - CRM and TFBS prediction: using PATSER to scan sequences of each species (independently) and CRM as cluster of TFBSs. Identified 18 enhancers in eve locus: stripe enhancers and muscle-heart enhancer (MHE), and ortholog with Dmel enhancers can be defined with the 20-30 bp blocks.
- Results:
  - Sequence comparison: very low identity of non-coding sequences between sepsids and Drosophila, in the range of <10%. Furthermore, the binding site organizations are not conserved between the two families. However, there are blocks (20-30bp) of sequences that are shared between the two families.
  - Expression pattern of TFs: Hb, Kr and Gt very conserved, not able to examine Bcd, Cad and Kni. Eve pattern is vrey similar.
  - Expression pattern of enhancer sequences in Dmel embryos: eve 2, eve 3+7, eve 4+6 and MHE, all show conserved expression patterns in the two families. Exception: eve 3+7 enhancer show posterior expansion at stripe 7.
  - Conservation of paired TFBSs: if a TFBS is paired with another TFBS of AP factors (overlapped or within 10bps), it is much more likely to be extremely (in sepsids and Drosophila) or highly (in 12 Drosophila species) conserved.
- Discussion:
  - The increased conservation of adjacent paired sites: may be explained by (1) mutation affecting one site may affect the interacting pairs (cooperativity, synergy, SRR); (2) large deletions tend to remove adjacent sites (a larger fitness loss), and will be purified more often than those deletions that only affect individual sites.
  - Explanation of eve 7 stripe change: (1) more diffused control of eve 7 pattern in the regulatory region; (2) different trans- environment.

Evolution of NEE enhancers in Drosophilia [Crocker & Erives, PLoS Biol, 2008]

- Problem: are the expression patterns of CRMs under adapative changes, or stabilizing selection, in different Drosophila species? And what are the changes at the cis- level (ex. compensatory change for stabilizing selection; or site organization change for adapative selection)?
- Background:
  - Neurogenic ectoderm enhancers (NEE): of genes vnd, rho, brk, vn and sog.
  - The regulatory mechanism NEEs: low affinity binding sites of Dorsal (activator), and suppressed in the ventral side by snail (repressor).
  - The common organization of NEEs: one or two pairs of Dorsal-Twist sites (< 20 bp); and overlapped Su(H)-Dorsal sites.
- Results:

  - Endogenous expression of orthologous genes are very similar in D. pseudoobscura (Dp), D. melanogaster (Dm) and D. virilis (Dv).
  - The expression patterns of orthologous NEEs in transgenic fly of Dm are different: orthologous NEEs of Dp generally have narrower pattern than Dm ones, while those of Dv generally have broader patterns than Dm ones.
  - Changes at cis- level: in both brk and vn NEEs: the spacing of Dorsal and Twist sites is different in Dp and in Dv. If changing spacing in Dm NEE to mimic Dp (or Dv) NEE, the resulting expression pattern of mutated Dm NEE will be similar to Dp (or Dv) NEE.
  - Changes at trans- level:
    - Change of Dorsal protein sequence: Dorsal mutations in Dp lead to increased affinity, and those in Dv lead to reduced affinity.
    - Expression of Dorsal: slightly narrower expression in Dv and broader expression in Dv.
    - Availability of Twist: more Twist sites in the genomic background in Dv, thus reduce the Twist availability.

- Discussion: the basic hypothesis is: changes at the trans- level, perhaps including both the Dorsal affinity and Dorsal expression pattern are compensated by the changes at the cis- level, mainly through tuning the distance (linkage) between Dorsal and Twist sites. Specifically,
  - In Dp: increased Dorsal affinity and broader Dorsal expression $\Rightarrow$ weaker NEEs $\Rightarrow$ narrower expression in Dm;
  - In Dv: reduced Dorsal affinity and narrower Dorsal expression $\Rightarrow$ stronger NEEs $\Rightarrow$ broader expression in Dm.

- Remark:
  - The trans- level changes can be important sources of regulatory evolution, driving the changes of cis-regulatory sequences.
  - The change of linkage between two sites is one type of cis- change, other changes may also be important, e.g. the number and affinity of sites. Ex. sog NEE of Dv contains fewer Dorsal sites (this is inconsistent with the changes at trans- level).

Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome [Ni & White, PLoS Biol, 2012]

- Background: insulator proteins function as barrier against heterochromatin spread, regulate promoter-enhancer communication by preventing in appropriate interactions. CTCF is the only known insulator conserved between human and fly. It is crucial in epigenetic imprinting, X inactivation.

- Data: ChIP-seq of three replicates (each has input) in four Drosophila species.

- Peak calling in ChIP-seq: use QuEST to obtain the CDP scores, then to call peak in one species: (1) normalization of CDP in each sample (IP or control) by the read depth; (2) mean CDP enrichment score: CDP scores in IP - CPD score in control, then take the average. Use permutation to determine the threshold of mean CDP enrichment score. To call peaks, further add the fitler that the fold enrichment (ratio of mean CDP in IP over mean CDP in control) must be ≥ 2.

- Comparison of binding profiles across species: call peaks then comparison is not optimal. So first map all the reads in non-Dmel to Dmel, and normalize the count of reads by the difference of reads and genome size. Then obtain the normalized CDP scores, and model them as a linear model:

$$Y = B_E \cdot \text{Experiment} + B_S \cdot \text{Species} + B_I \cdot (\text{Exp} \times \text{Species}) + \epsilon \qquad (3.2)$$

Then we test if B_E, B_S (species-specific enrichment) and B_I (species and treatment interaction) are equal to 0, and obtain p-values. To decide if a site is conserved, it must satisfy: the species enrichment scores in both species must be significant.

  - Related work: use $p < 10^{-21}$ as the threshold for Dmel, and a relaxed threshold $p < 10^{-5}$ for non-Dmel species to define conserved sites.

- Rate of CTCF binding divergence: about 2,000 and 3,000 peaks in each species. To study divergence, distinguish two scenarios: divergence due to sequence change; and divergence of binding only. Overall, found that diverge substantially, e.g. Dmel-Dpse comparison, 68% of CTCF sites in Dmel in fully aligned regions have diverged in Dpse. Average: 2.22% per Myr, compariing with 1.19% for protein sequence and 6.34% for synonymous sites.

- Molecular mechanisms of CTCF evolution: correlation with sequence evolution. Study in two ways: (1) conserved CTCF sites are more conserved at sequence level. (2) Conserved CTCF sites more often have conserved CTCF motifs. Also new CTCF sites: transposable elements (TE) do not play a major role.

- Role of natural selection in CTCF site evolution: study the polymorphis data of 37 Dmel inbred lines to infer selection on CTCF sites (defined as flanking 200bp region, or motif matches)
  - Negative selection: estimate Tajima's D in CTCF sites. More negative selection in CTCF sites than synonymous sequences, comparable to 5' and 3' UTR.
  - Positive selection: MK test to estimate α, the proportion that is fixed. Higher α in diverged CTCF sites than conserved ones, suggesting positive selection.

- Adapitve force that drives positive selection:
  - Expression evolution: determine stable or divegered genes by expression. Diverged CTCF sites are more likely to be near diverged genes than conserved CTCF sites.
  - New genes: 42 new Dmel genes, 8 have new CTCF sites.

- Lessons:
  - Comparison of binding (or other genomic properties) across species, the technical problems are: cannot discretize then compare; some part of the genome may not have orthologs.

3.5 Regulatory Evolution in Vertebrates

Regulatory evolution and human uniqueness [Sholtis & Noonan, TIG, 2010]

- Goal: characterizing putative CREs in human genome.

- Conserved noncoding sequences (CNS):
  - CNSs tend to be disproportionately located near genes with developmental functions.
  - Many sequences that fail to show significant conservation may function as enhancers in vivo.

- Human-accelerated CNS (HACNS):
  - Possibe artifacts of identifying HACNS: fast-evolving regions of the genome (local rate), alignment.
  - Biased gene conversion (BGC): a recombination-driven process that increases the fixation probability of AT to GC mutations, producing an increase in the local background neutral substitution rate. BGC can be distinguished from HACNS by: BGC substitutions tend to skewed towards GC; often in the region of high recombination rate; often extend beyond functional elements.

  - Does BGC produce false postives of HACNS? Perhaps limited effect, control of local substitution rate, and BGC often deleterious or no effect, while several HACNS studied to data do show gain-of-function.
  - Estabilishing adaptive values of CREs: several levels of proof: (1) Molecular changes: gene expression; (2) Reverse genetics: mouse carrying the human version of CREs show human-like phenotype; (3) The trait is associated with changes generally considered to be adaptive at some point in human evolution.

Conservation of upstream sequences in human/mouse [Iwama & Gojobori, PNAS, 2004]

- Idea: the regulatory constraints on genes are different, some class of genes may under more constraint.
- Methods:
  - Use upstream 8kb sequence of human and mouse orthologs (presumably enriched with CRE)
  - Alignment: local alignment (BLAST2) + post-processing: e.g. remove overlapping local alignments → global pairwise alignment
  - Quantify conservation for each gene: number of identical nts. in the alignments
  - Study representation bias of gene categories in highly-conserved genes: testing use binomial distribution
- Results:
  - Among the highly-conserved genes, TF's are overrepresented. Ex. ZFHX1B (homoeobox): 6,000 identical sites.
  - Developmental TFs show higher conservation of upstream sequences than other TFs
  - Lack of repeats (transoposons) in the upstream sequences of the top genes.

Suppression of long indels in CRM [Cameron & Davidson, PNAS, 2005]

- Methods:
  - Species: two closely related sea urchin (18 Myr)
  - Data: 5 CRMs (3 TFs and 2 signaling ligands) + endo 16 vs flanking sequences (the species are sufficient close s.t. even the flanking sequences can be aligned)
- Results:
  - Almost complete suppression of long indels (¿20bp); substitutions and small indels are 30-50% lower, which could be due to the constraint inside or immedidately adjacent to the binding sites.
  - Almost no single-base changes and small indels within the known important binding sites
- Discussion/Conclusion: In the flanking sequences, long indels will quickly destroy alignment; within modules, the suppression of long indels and the constraints on the sites themselves allow the sequences to be still alignable (thus appear conserved)
- Discussion: species divergence effect (on predicting CRM)
  - If too chose, not enough change in both CRM and flanking, thus undistinguishable
  - If too distant, only a small fraction (TFBS) will be conserved → no patchy pattern of conservation
  - Over relatively short distance, CRM still alignable (because of suppression of large indels + conserved TFBS), but not for flanking regions

Nature selection on human miRNA sites [Chen & Rajewsky, NG, 2006]

- Problem: are microRNA (miRNA) sites under natural selection? And what fraction?
- Background: miRNA, once expressed, could inhibit genes through binding of 3' UTR miRNA-binding sites in the target gene mRNA.
- Methods:
  - Data: 25,000 human SNPs, and 22,000 predicted miRNA sites conserved in 5 mammals
  - SNP density at different regions/classes of elements: low SNP density suggests negative selection
  - DAF distribution of SNPs: higher fraction of rare SNPs suggests negative selection
  - M-K test
  - PRF analysis of polymorphism spectrum to estimate the selection. In particular, a mixture of sites where some sites are under negative selection and others under positive selection or neutral
- Results:
  - SNP density is lower in conserved miRNA sites than conserved K-mers; SNP density also shows positional variation: lower density in presumably more important positions (leading positions in the binding sites)
  - rare SNPs are more common in conserved miRNA sites than in other classes of elements: nonsynonymous sites, conserved 7-mers, etc.
  - M-K test not significant, possibly because positive selection reduces the signal of negative selection on miRNA sites
  - PRF mixture analysis of conserved miRNA sites: about 85% are under selection; the same analysis of miRNA sites that are coexpressed with mRNA (the miRNA genes and the target genes are expressed in the same tissue): about $30 - 50\%$ are under selection.

Beak morphology in Darwin's finches [Abzhanov & Tabin, Nature, 2006]

- Problem: genetic basis of the evolution of beak morphology (length)?
- Results:
  - Correlation between CaM expression and beak length in different species suggests CaM as a candidate gene: (i) validity of transcriptional profiles: the expression profile tree agrees with the species tree; (ii) CaM is highly expressed in finches with long beak: both by microarray experient and by in situ hybridization
  - constitutively active CaMKII (CaM kinase kinase, a downstream target of CaM) in chicken leads to the increase of beak length
- Discussion: independence of beak length (by CaM) and width/depth (by BMP4)
- Remark: the same principle of associating gene expression with phenotype (to identify candidate genes) also applies in the study of evo-devo: where phenotypes are in different species

Divergence of TF binding in liver cells of human and mouse [Odom & Frankel, NG, 2006]

- Problem: is transcriptional regulation (indicated by binding of key TFs) conserved across human and mouse?
- Methods: ChIP-chip of 4 liver TFs: FOXA2, HNF1A, HNF4A, HNF6, in human and mouse. Promoter sequences (up/down 5k) of 8000 orthologs.
- Results:

  - Substantial divergence: 41%-89% of promoters bound by a protein in one species were not bound by the same protein in the second species. Conserved binding in aligned region generally accounts for only about 1/3 of binding.
  - Not due to intra-species variation: compare different human liver cells (one transformed line).
  - Not due to TF binding specificity divergence: almost identical PWMs.

Expression profiling in primates reveals a rapid evolution of human transcription factors [Gilad & White, Nature, 2006]

- Motivation: use inter-species comparison of transcriptome to study the pattern of selection: how often and what are the strongly selected genes.
- Data: multi-species array of liver samples from human, chimp, organtans and rhesus. This array fixes the issue with sequence hybridzation using human array.
- Method of testing selection: linear mixed model treating species effects as fixed effect and indivudal variation as random effect. The genes with the same mean are candidates of negative selection; and genes with lineage-specific mean as candidates of positive selection.
- Negative selection of gene expression: a majority of genes (60%) do not show significant inter-species expression difference. This could result from true stabilizing selection or large intra-species variation. Ranking of these genes by low intra-species variation to find candidates of negative seleciton. The top genes are enriched with transcriptional regulation. The results support that negative selection is the dominant mode.
  - Remark: previous results suggest that expression evolution is neutral, based on evidence that expression divergence is proportional to intra-species variation.
- Positive selection of gene expression: identify 14 genes with significantly higher expression in human and 5 with lower expression. Among genes with higher expression, 5/12 are TFs. Correspondence of expression and sequence evolution: genes whose expression are under (+) selection are more likely to have (+) selection in coding sequences (25%); while the ratios for genes that are under (-) selection or no selection are much smaller: 6% and 4% respectively

Human and rodent male gametogenesis [Chalmel & Priming, PNAS, 2007]

- Goal: find tissue-specific (testicular germ cells) genes, i.e. gene specifically expressed in a certain tissue type.
- DET: differentially expressed in testis. 4 types of cells in testis - somatic (SO), mitotic (MI), meiotic (ME), and post-meiotic (PM)
  - Differential expression: significantly different values in the 4 conditions
  - Clustering: s.t. one cluster has expression specific to one type of condition ⇒ 4 clusters corresponding to 4 types
  - 7066, 5140, 5119 genes respectively in mouse, rat and human
  - 7066 mouse genes form 4 clusters, each induced in a different cell type (SO, MI, ME, PM)
- CDET: conserved differentially expressed in testis
  - Genes that belong to the same cluster in all three organisms
  - Genes whose pairwise expression correlation (mouse and rat) is larger than 0.80
  - 1001 common genes that are differentially expressed in all 3 organisms
  - 888 genes are conserved and differentially expressed

- CDEST: conserved, differentially expressed and specific to testis
  - Methods: add 17 other somatic control (heart, eye, brain, etc.), out of 888 genes, which are not expressed in all 17 control cells
  - 80 genes - somatic, mitotic, meitoic (42), postmeiotic (33)
- Remark:
  - Clustering of experiments (tissue types, etc.): allow grouping of cell types or conditions, which can be used for analysis of gene expression pattern (choose two groups for differentiall expression, choose number of clusters, etc.)
  - Guided clustering of expression profile: choose number and meaning of clusters based on biology - typically, one cluster of genes corresponding to phenotype/cell type (P)represents genes specificall expressed in P
  - Appropriate selection of samples and controls is important: for example, to find testis specific genes, use various types of somatic cells as controls (instead of a single one)

p53 regulatory evolution [Jegga & Resnick, PNAS, 2008]

- Motivation: p53-target relations conserved across species? And why?
- Methods: Human-mouse: known targets (with known TFBS or predicted TFBS, called RE) and known activity (called transactivation potential) of each RE.
- Results:
  - Essentially no difference between sequence conservation and function conservation; less functional conservation than sequence conservation. For example: high conservation of sequence $\rightarrow$ different levels of function
  - Change (neither sequence nor function is conserved) in DNA metabolism and repair genes. Likely due to change of life sytles of rodents (nocturnal, less UV exposure)

Constraints in primate regulatory regions [Gaffney & Majewski, PLoS Genetics, 2008]

- Aim: what are the factors for selective constraints of regulatory sequences: the species (population size), expression pattern of target genes, etc.?
- Methods:
  - Data: primates (i) experimental TFBSs from TRANSFAC; (ii) grouping of TFBSs into pCRMs; (iii) ChIP-chip regions.
  - Control sequences (for defining constraints): divide all positive sequences into "case" regions, then for each case region, its control sequence is the intronic sequences (removing first intron) in the 500kb flanking region.
  - Selective constraint: $1 - O/E$ where O is the number of substitutions in sequences of interest, and E is the expected number in neutral sequences (parsimony: no multiple-hit correction). Linearge-specific rates were estimated with parsimony.
  - Expression breadth: the number of tissues a target gene is expressed.
- Results:
  - Constraints in TFBSs, pCRMs and ChIP regions: about 37% mutations are deleterious.
  - Lineage-specific constraints: reduced in human and primate vs macaques.

  - Constraints are negatively correlated with expression breadth: simple expression pattern (e.g. housekeeping genes) more likely to tolerate the mutations in CRS.

Evoulution of mammalian TFBSs via transposable elements [Bourque & Liu, GR, 2008]

- Methods:
  1. Data: ChIP dataset of 7 mammalian TFs: ESR1, TP53, MYC, RELA, POU5F1-SOX2 and CTCF
  2. Measuring conservation of binding regions: (i) percentage of binding regions that overlap with conserved elements from PhastCons; (ii) percentage of binding regions that contain a conserved binding site in human and another mammalian homologous region.
- Conservation of ChIP bound regions: (i) 10-40% regions overlap with PhastCons conserved elements; higher, but to a limited degree (no more than 2-fold) than the random control; (ii) similarly 20-30% regions contain conserved binding sites; much higher than random control → Overall binding regions are under more constraint (and if measured by binding site occurrance, much higher) than random sequences, but the absolutely level of conservation is low.
- Dependence of conservation on positions: divide bound regions into 4 groups - adjacent (within 250 bp of TSS), proximal (5kb), distal (100kb) and desert (> 100 kb). The percentage of bound regions that contain conserved binding sites: in Myc, higher in adjacent and proximal ($30 - 40\%$) than in distal and desert ($< 10\%$); in ESR1 and CTCF, similar ($30 - 40\%$ in all groups).
- **Remark:** a significant fraction (not necessarily majority) of regulatory sequences is under evolutionary constraint, which makes these sequences overall more conserved than random sequences (but the constraint itself is not homogenous).

Evolutionary constraints of GATA1 binding sites [Cheng & Hardison, GR, 2008]

- Aim: the evolutionary constraints and biological activities of binding sites.
- Methods:
  - Data: GATA1 ChIP-chip on mouse chromosome 7.
  - Function of GATA1 bound region: regulatory activity in erythroid cells.
- Conservation of GATA-1 binding sites: almost all of the occupied segments contain canonical WGATAR motif, in only 45% of the cases is the motif deeply preserved.
- GATA-1-bound segments with high enhancer activity tend to be the ones with an evolutionarily preserved WGATAR motif.

Adaptive changes of HoxA-11 [Lynch & Wagner, PNAS, 2008]

- Problem: prolactin (PRL) expression in endometrial cells is a novel expression pattern of PRL in placental mammals. How was it evolved?
- Background: HoxA-11 plays multiple functions in blood cell differentiation, the development of body axis, limbs, kidney and male and female reproductive systems.
- Methods:
  - Test HoxA-11 function: PRL expression in endometrial cells. Different versions of HoxA-11 can be tested: placental, non-placental, ancestral.
- Evolution of PRL enhancer: the insertion of MER20 element upstream of PRL, 166-175 MYA (coincident with a period of rapid evolution in HoxA-11).

- Evolution of HoxA-11 protein:
  - PAML analysis: 10 sites under positive selection (posterior probability > 0.90 under BEB, according to branch-site test) in the stem-lineage of placental mammals. They are under extremely strong negative selection within the extant placental mammals ($\omega = 0.08$ vs $\omega = 0.44$ in non-placental mammals).
  - DIVERGE analysis: compare two clades: mammals and birds/reptiles. $\theta_I = 0.46 \pm 0.18$, $\theta_{II} = 0.09 \pm 0.05$. This suggests that selection recruited weakly constrained AAs into a novel function rather than AAs that were under strong preexisting functional constraints.
- Only HoxA-11 in placental mammals, but not in nonplacental mammals or ancestral, is able to activate PRL expression.

Conservation of gene expression in vertebrate tissues [Chan & Hughes, J Biol, 2009]

- Problem: are gene expression patterns in tissues conserved across vertebrates? And if so, are the non-coding sequences conserved accordingly?
- Idea: measure the conservation of gene expression, and test if it is correlated with the sequence evolution.
- Data: expression of orthologous genes (3,074 genes) in 10 major tissues in human, mouse, chicken, frog and pufferfish.
- Conservation of gene expression: measured in two ways:
  - Binary measure: expression in each tissue is discretized: 1 or 0, by using some cutoff (ranks of expression level among all genes, top 1/2, top 1/3, top 1/4, et.c). A gene is conserved in one tissue if it is expressed in this tissue across all 5 species.
  - Pearson correlation: define the expression profile of a gene across 10 tissues, and pairwise conservation of a gene is measured by the Pearson correlation of the two expression profiles of this gene in the two species.
- Measuring the conservation of non-coding sequences of a gene: choose 50 kb upstream and downstream of a gene. The conservation is measured by: the number of bps (or the proportion) in conserved elements found by PhastCons; the number of conserved elements found by a method that allows non-local alignment; EEL scores using 138 JASPAR motifs.
- Tissue expression profiles are broadly conserved across species: the profiles of a tissue in different species are more similar than the profiles of unrelated tissues (in the same species). Importantly, some tissues are more conserved than the other ones: kidney and testis are the least conserved tissues.
- Many gene expression events are conserved across five species: about 20% of genes or gene expression events (a event is gene expression in a tissue) are conserved, measured in two different ways. 1,488 genes are conserved by at least one of the two measures.
- The exprssion conservation is not correlated to non-coding sequence conservation.

TFs in human-chimp expression comparison [Nowick & Stubbs, PNAS, 2009]

- Hypothesis: the different expression of TFs in brain in human vs chimp contribute to phenotypic difference.
- Data: [Khaitovich, Science05] five tissues (brain, heart, liver, kidney and testis) of 6 humans and 5 chimps.

- Define differentially expressed genes: must satisfy all critiria (1) $P < 0.05$ under t test, with multiple hypothesis correction; (2) have difference of at least 1.2 fold; (3) difference of at least 20 units of expression values (modest level of expression in at least one species).
- Define genes associated with TFs: by expression correlation in 30 human tissues (6 samples, 5 tissues each), also require that the humna-chimp brain expression difference must be in the same direction as the TF.
- 90 TFs show differential expression in brain in human vs chimp, including 33 KRAB-ZNFs. The tother TFs correspond to many protein families.
- Brain TF network: two TFs are linked according to expression correlation and the overlap of their associated genes. This network reveals two modules: Module 1 - human up-regulated; Module 2- human down-regulated.
- Functions associated with modules: (1) Module 1: transcription, vesicle transport (neurite outgrowth, axonal transport and synaptic transmission), ubiquitination and unfolded protein response (neuroprotective functions); (2) Module 2: energy metabolism including ROS metabolizing function.

Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding [Schmidt & Odom, Science, 2010]

- Data: CEBPA and HNF4A ChIP-Seq in human, mouse, dog, opossum and chicken. Both TFs are conserved (DBD nearly identical across 5 species) and constitutively expressed.
- Binding profiles: less than a quarter of bound regions within 3kb of TSS, and between 30 to 50% of the binding sites of the two TFs overlap.
- High divergence of binding: (Figure 2)
  - In pairwise comparison between mammals, only 10 to 22 % binding events are shared (in aligned regions). The ratio is only slightly higher when alignment is adjusted.
  - Even higher divergence in other vertebrates: e.g. opossum, only 6-8% binding of opossum occur in aligned regions in mouse, dog and/or human liver; in chicken, only 2% of CEBPA binding is shared with human.
- Functional targets: 35 binding events are conserved across all 5 species, near geens central to liver biology; the target genes of the two TFs tend to be conservd in at least two species; functional liver enhancers (38 out of 53 enhancers in 9 HNF4A bound regions, out of which five overlap with CEBPA binding) tend to have shared binding.
- Binding specificities remain virtually identical.
- Turnover of binding sites:
  - The lost binding events are associated with binding site loss ($> 50\%$ cases), and similar for gains.
  - In about half of lineage-specific losses, a compensatory gain of binding events can be found within ±10 kb (Figure S16). Most of these compensatory gains occur within 2Kb of the orthologous sites.

Conserved expression without conserved regulatory sequence [Weirauch & Hughes, TIG, 2010]

- Goal: the conservation of expression with divergent sequences, observation and explanation/rationalization.
- Examples of expression conservation with divergent sequences:
  - Developmental enhancers: e.g. eve enhancer.

  - TF substitution/rewiring: Tbf1 to Rap1 in RP genes; Mcm1 in mating type; GalX (unknown) to Gal4 and Mig1 in galctose genes.

- Mechanism of cis-regulatory turnover:
  - Evolution of TFs: binding specificity generally change very slowly.
  - CRS turnover and shufflign: permitted by the billboard model.

- Benefits of tinking: cis-regulatory turnover and shuffling may be a by-product of organization schemes that ensure consistent while facilitating variation and neofunctionalization.

- Future directions:
  - Sequence-to-function mapping of CRS: more data from epigenetics, TF-binding specifities, etc.

A comparative encyclopedia of DNA elements in the mouse genome [Mouse ENCODE Consortium, Nature, 2014]

- Data:
  - 123 mouse cell types and tissues (http://www.mouseencode.org/data).
  - Brain: histone marks (H3K27ac, H3K4me1, etc.), RNA-seq, DNase-seq in whole brain, Cerebellum, Cortex, frontal lobe.
  - ChIP-seq of 37 TFs in various subsets of subsets of 33 cell/tissue types.
  - Histone marks: use a random forest based classifier to predict CREs from H3K4me1, H3K4me3 and H3K27ac. Valiadation rates: 87 and 71% for predicted promoters and enhancers, respectively.

- Conservation of TF networks:
  - Only 22% TF footprints are conserved, but nearly 50% of cross-regulatory connections (TF-TF) are conserved between mouse and human.

- Comparison of transcriptome:
  - Initial analysis: gene expression patterns tended to cluster more by species rather than by tissue
  - About 4,800 genes drive clustering by species, rather than tissues. Variance component analysis: partition the variance by tissues or species. Removing them reveals tissue-species expression.
  - NACC analysis: assess the conserved co-expression. For each gene, ask if the co-expressed genes in one species are also co-expressed in another species. Specfically, for a gene (g) in human, find its 20 closest neighbors; then quantify the average distance between g and the 20 genes in mouse. Do it then for mouse-human comparison. The average distance between the two comparison is the NACC score for the gene.
  - Genes highly conserved are: transcription and regulation, intracellular metabolic processes. Highly divergent: immune, extracellular matrix, etc.

- Comparison of CREs:
  - Homologous sequences: 56.4% of the enhancer predictions, 62.4% of promoter predictions, 61.5% of DHS, and 53.3% of the transcription factor binding sites have homologues (50% bp alignment), compared with an expected frequency of 34%, 33.8%, 33.6% and 33.7% by random chance.
  - Putative lineage-specific CREs: 15-18% of candidate mouse promoter/enhancers have no homolog in human, the majority of which show function in experiments.
  - Overlap with repetitive sequences: 85-89% of mouse-specific promoters/enhancers overlap with repetitive sequences (compared to 78% by random chance).

  - Conservation of function of homologous CREs: of mouse predicted sequences, 44% (22,655) are still predicted as promoters in human and 40% (64,962) of them are predicted as an enhancer. Further evidence from NACC analysis on promoter/enhancers (higher conservation than expected).

- Remark:

  - Advantage of NACC is: no need of exactly matched samples.

Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. [Vierstra & Stam, Science, 2014]

- Data: DHSs in 45 mouse cell and tissue types: adult primary tissues and cells (n=29), primary embryonic tissues (n = 4), embryonic stem cell lines (n = 4), and model immortalized primary (n = 3) and malignant (n = 5) cell lines.

- Conservation of mouse DHS: 60% of mouse DHS can be aligned to human genome, of which 35.6% coincide with one human DHS.

- Role of transposons in evolving novel regulatory elements:

  - OCT4 (and other POU family TFs) binding sites have been greatly expanded in mouseon a LTR/ERVL element, accounting for $> 25\%$ of mouse-specific sites versus $< 5\%$ in humans.
  - Expansions of CTCF sites were largely driven by short interspersed elements (SINEs) in both mouse and human.

- Tissue selectivity of shared DHS:

  - The vast majority of shared DHSs (78.8%) displayed tissue-selective accessibility and were organized into distinct cohorts (Figure 2A). A cluster represent a broad tissue origin, e.g. ESC, brain, retina, lung, heart, fibroblast, T cells, B cells.
  - A minority (21.2%) exhibited high accessibility across multiple tissue types, whereas $< 5\%$ were constitutive
  - Tissue-selective DHSs showed pronounced enrichment for nearly all known lineage-specifying regulators. Ex. Oct4, Sox4, Klf4 in ESC, Klf4 in intestine- and erythroid-specific DHSs (KLF-family TFs share similar motifs)

- Repurposing of DHS between mouse and human: Among all shared DHS, substantial repurposing (active in a different cell type) ranging from 22.9 to 69% of shared DHSs, depending on the tissue. Overall, at least 35.7% of shared DHSs were repurposed.

  - Highest conservation: brain, retina, muscle, intistine
  - Lowest conservation: liver, erythroid (red blood cell)

- Conservation of TF binding motifs within shared DHS:

  - 39% of TF recognition sequences were positionally conserved and 20% were operationally conserved.
  - 41.3% of shared DHSs (chiefly repurposed DHSs) lacked any positionally or operationally conserved TF recognition elements.
  - The overall density of TF recognition elements did not differ substantially between shared DHSs with positionally, operationally, or nonconserved TFs
  - Recognition sites for cell fate?modifying TFs were consistently depleted within repurposed DHSs.

  - Striking conservation of the proportion of the regulatory DNA landscape of each cell type devoted to recognition sites of each TF: the role of each TF in a given cell type is conserved, and regulates a similar number of DHS.

- **Summary**:
  - Individual DHSs are not highly conserved: 36%; and individual TF recognition sites are even less conserved, about 22%.
  - Even shared DHS may change activitiy (tissue-activity profile), about 40%. The mechanism is likely the change of TF recognition sequences. This may be one way strategy of evolving new functions: repurposing existing CREs by changing affinity to TFs.

Evolutionary changes in promoter and enhancer activity during human corticogenesis [Reilly & Noonan, Science, 2015]

- Background: humans exhibhit difference of corticogenesis within the first 12 weeks of gestation. The differences include: an increased duration of neurogenesis, increases in the number and diversity of progenitors, modification of neuronal migration, and introduction of new connections among functional areas.
- Data: H3K27ac and H3K4me2 to map active promoters and enhancers during human, rhesus macaque, and mouse corticogenesis.
  - In promoters: 85% have both marks. In enhancers, 45% by both.
  - Validation of data: PCA of samples, and using PCs, samples are clustered by cellular origin instead of organisms.
  - 9K nonoverlapping enhancers and 2.8K promoters show epigenetic gain in human
- Sequence change of enhancers/promoters with human gain: not show increased rate of human-specific change.
- Biological function of human-gain promoters/enhancers: map to nearest genes
  - Use co-expression networks defined by BrainSpan data: 96 modules.
  - 17 modules are enriched for human gain epigenetic changes.
  - Module 3: enriched with neuronal progenitor proliferation, associated with genes important for cortical development such as PAX6, GLI3, FGFR1
  - Module 10: extra-cellular matrix (ECM), contributes to maintenance of human progenitor cell self-renewal and neuronal migration.
- TF/motif analysis: predict motifs enriched in enhancers or promoters assigned to each module. Many motifs were enriched in promoters and enhancers assigned to the same module as the TF itself.
  - 591 transcription factor binding motifs from JASPAR (vertebrate core motifs) and Uniprobe.
  - All three-way orthologous H3K27ac and H3K4me2 regions in human (including gains) were scanned for the presence of each motif using FIMO.
  - Permutation test: count the number of motif instances in all promoters or enhancers assigned to each module. Then in each permutation, we obtain the number of motif instances called in an equally sized set of enhancers or promoters derived from all three- way orthologous regions.
- Questions:
  - Human-gain: changes in underlying sequenes (cis), or in TFs (trans)? Ex. analysis of BSs of the TFs identifed. Possible that TF change leads to change of epigenetic states of many enhancers without any cis-changes.
  - Signatures of negative selection in human population?

3.6 Other Cases of Regulatory Evolution

1. Module conservation [Balmer, Biology Lett, 2006]

   Methods:

   (a) Species: human (Hs), mouse (Mm), rat (Rn)

   (b) Data: 77 known binding sites of retinoic acid receptor (RAR, a nuclear receptor) merged from all 3 species. Many types of target genes: Hox cluster, ATP-binding cassette transporters, regulators of glucose metabolism, etc.

   Results:

   (a) In all but 5 cases, the homolougous region surrounding the TFBS exist

   (b) Median module (neighborhood homologous regions) length is 500-600. High percent identity in the module: 86% in Mm and Rn, 63% between Hs and Mm.

   (c) Median number of blocks ¿= 5 nts = 15.5

2. Aging in worm and fly [McCarroll & Li, NG, 2004]

   Problem: the transcriptional profile of aging, how it is conserved or diverged across species

   Methods:

   (a) Data: (i) C elegans: cDNA microarray; young vs old → aging profile; (ii) D mel: oligo-nt. arrray; young vs old → aging profile.

   (b) Global profile comparison: through Pearson correlation (r)

   Results:

   (a) Aging process has a conserved module across species. Test: correlation between worm and fly aging profiles

   (b) Genes and gene groups involved in aging process are conserved across species. Test: conservation of GO-defined gene groups

   (c) General method of computing conservation of genomic expression pattern: via r. Find the most similar global expression profiles in a database of profiles

3.7 Regulatory Evolution: Mathematical Models

1. Simulation study of TFBS gain [Stone & Wray, MBE, 2001]

   Model:

   (a) Promoter sequence of length 200-2000bp under local point mutations

   (b) Each sequence is doing mutational random walk with rate = 10^{-9} per bp per generation and if the effective population size is N, then N sequences are doing mutational random walk independently

   (c) The waiting time for a new TFBS is determined by: the waiting time for a single sequence divided by N. Only exact match is considered.

   Results: suppose $N = 10^6$

   (a) 6-bp TFBS: $4.5 \cdot 10^9$ (for single sequence) divided by $(2 \cdot 10^6) = 2254$ generations

   (b) eve-stripe 2: in about 600bp, takes 65 years to fix a 6-bp binding site.

   Criticism:

(a) The population size is incorporated incorrectly: not independent!

(b) Fixation of BS is not considered (the appearance of a new TFBS via mutational random walk will be fixed with a small probability).

(c) Seriously underestimate the waiting time

2. Evolution of pairs of compensatory mutations [Carter & Wagner, PRSLB, 2002]

   Problem: enhancers are much more conserved in mammals than in Drosophila. Could it be due to differences of population size and generation time?

   Idea: a pair of individually deleterious, but compensating mutations, are fixed faster in large populations.

   Model: consider an allele A, it has two deleterious mutations ($1+s_d, s_d < 0$), but when both mutations happen, it is slightly advantegous ($1+s$). The evolution from w.t. A to double mutants may follow two pathways:

   - Fixation of one mutant, then fixation of the other. The time of fixation (t_1) is determined from two mutation and fixation events.
   - Before the first mutant is fixed, a second mutation in some individual with the first mutant, thus creating the double mutant, which will be fixed. The time of fixation (t_2) is determined from (i) the time that the second mutation occurs while the first one is still not fixed; (ii) the time of fixation of the allele with both mutations.

   The overall rate of fixation of the pair is given by:

$$R = \frac{1}{t_1} + \frac{1}{t_2} \tag{3.3}$$

   Methods:

   (a) Simulation procedure: N diploid individuals, at each generation, each allele has a probability of mutation (non-reversible). After this step, N pairs of individuals were picked (sequentially with replacement and probabilities weighted by their fitness), and mated to produce the members of the next generation.

   Results: under the setting, $\mu = 10^{-5}, s_d = -0.01, s = 0.001$, the dependence of rate on the effective population size is highly nonlinear (Figure 3). The rate is much higher for $N = 10^5 - 10^6$ (the invertebrates), than for $N = 10^3 - 10^4$ (vertebrates).

   Discussion: for small populations, pathway 1 dominates (faster fixation due to random drift); but for large populations, pathway 2 dominates as there will be more deleterious alleles that can segregate in the populations; as the population gets even larger, the rate is slower simply because of the increased time of fixation.

3. Position specific rate variation [Moses & Eisen, BMC Evol Biol, 2003]

   Aim: the evolutionary pattern of TFBSs.

   Methods:

   (a) Data: yeast promoter sequences; known TFBSs of Gal4, Gcn4, Mcm1, Abf1 and Rap1.

   (b) Prediction of BSs of other TFs: MEME of motif as well as binding sites on the promoter sequences of the known targets of a given TF (from microarray experiments).

   (c) Measuring evolutionary rates: substitution rates in multiple alignment - defined as the parsimony cost in the tree (only ungapped positions). Background evolutionary rates: from the whole promoter sequences.

Results:

(a) Known TFBSs of some factors: positional variation of rates corrrelate the information content. And the rates agree with the HB model.

(b) Predicted TFBSs of many factors show similar pattern (Table 2).

4. Expression evolution of gene families [Gu, Genetics, 2004]

   Problem: model of expression evolution.

   Model:

   (a) Basic model: the ancetral expression follows normal distribution $X_0 \sim N(\mu, \rho^2)$. The conditional distribution given ancestral expression follows: $X|x_0 \sim N(x_0, \sigma^2 t)$.

   (b) Lineage-specific model: σ^2 is lineage-specific

   (c) Directional trend: $X|x_0 \sim N(x_0 + \lambda t, \sigma^2 t)$

   (d) Punctuated model: change only at the begnning of each duplication

   Inference:

   (a) Likelihood function: for basic model, the likelihood function is obtained by integrating the ancestral expression:

   $$P(x_1, x_2) = \int P(x_1, x_2|x_0)\pi(x_0) = N(\mu, \mathbf{V}) \tag{3.4}$$

   where $\mu = (\mu, \mu)$ and

   $$\mathbf{V} = \begin{bmatrix} \rho^2 + \sigma^2 t & \rho^2 \\ \rho^2 & \rho^2 + \sigma^2 t \end{bmatrix} \tag{3.5}$$

   (b) Multiple experimental condition: iid assumptions, different ancestral level, i.e. treat them as evolutionary replicates

5. Simulation study of TFBS gain [MacArthur & Brookfield, MBE, 2004]

   Problem: how long does it take to evolve an enhancer sequence?

   Model: CRM as evolutionary unit

   (a) Fitness model: the CRM contains multiple BS, determined via cutoff of matrix similarity. The output of CRM is determined by:

      - The sum of output of each TFBS (its matrix similarity)
      - Position effect: favor TFBS clustering.

      Finally, the selection coefficient is proportional to its output.

   (b) Evolution: single genotype all the time. The substitution rate of one genotype to another is: $2N \cdot \mu \cdot$ prob_fixation, where prob_fixation is a function of the fitness of the two genotypes. Specifically, for any mutation, define the rate of change (in unit of μ, the mutation rate), S as:

      - If it is neutral, $S = 1/3$ (divide by 3 because μ is the total mutation rate per bp).
      - If it is advantageous, $S = 4Ns/3$, where N is the effective population size, and s is the selection coeffecient (the probability of fixation for advantageous mutation is approximately $2s$, thus the rate should be $2N\mu \cdot 2s = 4N\mu s$).

      Define S_{ij} as the i-th mutation (out of three) at the j-th position, then μS_{ij} is the rate of i-th mutation at the j-th position. The evolution will stop when an enhancer appears (the target output is reached).

   (c) Parameters: $N = 10^5, \mu = 10^{-9}$.

(d) Starting sequences: real Drosophilar enhancers and random sequences.

Results: the influence of various factors on the waiting time.

(a) Selection coefficient and starting sequences: increasing the selection coefficient (which is unknown) will lead to reduction of waiting time. The relation is linear for Kr730 enhancer, but not for random ones, suggesting selection drives evolution for Kr730 enhancer, but not random sequences. In other words, the existence of pre-sites in Kr730 enhancer (otherwise, the time will be spent on neutral random walk) greatly facilitates evolution.

(b) The effect of GC content, sequence simplicity, etc.: in general, it is easier to evolve TFBS whose content is similar to that of the sequence.

(c) TFBS length: shorter sites are not always easier to evolve, as they tend to be less flexible than longer ones.

(d) Selection threshold (when the sequence output is visible to selection) has a large impact on waiting time.

(e) Under stabilizing selection (maximum fitness when output > threshold): observed TFBS turnover.

Discussion:

(a) Single genotype assumption: only valid if time of successive fixations >> time of fixation. This is equilavent to:

- Neutral mutations: $4N\mu << 1$
- Advantageous mutations: $4N\mu_s \ln(4Ns) << 1$ where μ_s is the total mutation rate of advantageous mutations.

(b) Not possible to estimate a real waiting time, as that will need to know the selection coefficient, which is unavailable. Thus the main finding of this paper concerns the qualitative influence of pre-sites; the relative influence of other factors, e.g. the selection coefficient, the TFBS length (unexpected), etc.

(c) Haploids vs diploids: the simulation is based on $2N$ haploids. In the realistic diploid population, s should be seen as the selective advantage of the changed allele when it is in the heterozygous state, since this is what will determine the probability of its eventual fixation.

6. Rates of TFBS gain [Durrett & Schmidt, Annals of Applied Prob, 2007]

Problem: waiting time of a word W (also the length, $W = 6$ or 8)

Model:

(a) Neutral random walk: of a single sequence or a population. This is different from [MacArthur & Brookfield, MBE, 2004], which models adaptive evolution.

(b) Evolution of population: Moran model - continuous time birth-death process (birth: reproduction, probability depends on fitness; death: random so that the population size is constant).

Results:

(a) Single word of size W:

- Waiting time := time that the word under evolution becomes W
- Approximation 1. suppose mutation rate is 1 per nt., then waiting time is exponential distribution with mean $4^W/W$
- Analysis: consider the stochastics process X_t = number of matches to W at time t

(b) Segment of length L:

- Waiting time := time that W appears somewhere in the segment

- Approximation 2. (naive estimation) W = 8, exponential with mean $4^W/(LW)$.
- In initial segment, E(number ofmatches) = $L/4^W$

Remark: in both initial sequence and sequence during evolution, the expected number ofmatches is proportional to L.

(c) Population of words of size W under Moron process

- Waiting time := time that some word in the population becomes W
- Approximation 3. use the fixation chain Y_n, and let L_n = number ofmatches of Y_n. Then waiting time has mean $E(S)/(W\mu)$, where S is the time of fixation s.t. $L_n = W - 1$ and mu is the mutation rate/nt.

(d) Population of sequences of length L under Moron process

- Waiting time := time that some sequence in the population contains W
- Initial population: if match minus 1 in the population, then quick success. Because only one mutation is needed, rate = $(2N\mu)/3$.
- Otherwise, take much longer. Ex. $N = 10^4, W = 8, L = 1000, \mu = 10^{-8}$. If there exists match minus 1 in the initial population (prob = 0.3127): 375,000 yrs; o/w: 650 myr. Note: match in the initial population is defined via population consensus sequence (the majority of population).

Discussion:

(a) Importance of pre-sites: without them, the time of evolving a new site will be too large (650 myr for human, $W = 8$).

(b) Fixation probability: not included in this study, simply assuming that any new site will be fixed. Including it will further increase time.

(c) Large population size: sequences are more fluid, the exact treatment is not done here.

Extensions/Open problems:

(a) Energy-based instead of exact sequence based. One could reduce this to sequence/word based however.

(b) The case where $4N\mu \approx 1$ (or larger)

(c) For general equilibrium distribution and mutation rate matrix: e.g. when e.b. distribution is PWM and rate matrix is defined via HB model, and W is non-functional site, then the same method can be used to study TFBS loss.

7. Stationary TFBS turnover model [Wagner & Stadler, JTB, 2007]

   Methods:

   (a) Model: consider a genome region (in this study, relatively large region, HoxA cluster, > 100kb), BS origination follows a time-homogenous Poisson process with rate lambda and the death of each BS follows the exponential process with rate μ / site. The number of TFBS is the random variable, whose conditional distribution can be analytically determined: $P(X(t) = n|X(0) = x_0)$

   (b) Fitting the model with real data:

   - μ: the expected number of shared homologous sites should decrease exponentially with time since lineage separation, thus log. of the shared homologous sites is linear to divergence time, this would allow to determine the death rate through lineage regression
   - λ: the expected equilibrium number of BS is $n = \lambda/\mu$, thus $\lambda \approx \mu n$

   Results:

(a) Apply the model to the number of ERE (estrogen response element) HoxA cluster region, and obtain: $\mu = 1.3e - 8$ per yr, or, half life 27 Myr; $n = 12$; $\lambda = 1.6e - 7$ per yr.

(b) The log. of the number of ERE is roughly linear to divergence time, however, two outlier data point exist.

Remark:

(a) Birth and death rate independence of the current number of BS: valid only when the density of BS is very low. Otherwise, there will be compensation, redundency/competition, etc.

(b) Applications of the model: time heterogenity of the rates/change of selection over different lineages; difference of turnover rates of different TFBS or different regions.

8. PSPE [Huang & Ohler, Genome Biol, 2007]

Aim: simulate the evolution of cis-regulatory sequences.

Methods:

- Functional constraints of promoter sequences: TFBS distance to TSS (min and max); DNA strand; number of copies (min and max).
- Simulation: ancestral sequences - 3rd order Markov chain trained from a collection of human promoter sequences (or user-specified); substitution - HKY model; indel - Poission events and length follows geometric or negative binomial. The sequence evovles according to the substitution and indel models, subject to the functional constraints.
- Sequence homology vs functional homology: (Fig. 1) suppose the sequence requires (a, b, c) sites, then its descendent may be (a', b', c') where a', b' and c' are binding sites, but may not be exactly descendent of the ancestral sites (replacement at a different place).
- Replacement turnover rate (RTR): consider a TFBS in a given an ancestral sequence, simulate many times, the frequency the TFBS is replaced by another binding site but at a different place is RTR.
- Alignment benchmark: 6 binding sites with exactly one copy each (min = max = 1). The functional constraints also require them to be in the same order (via TSS distance constraints). Performance is measured by the TFBS detection accuracy, defined via functional homology (instead of sequence homology).

Questions:

- Simulation: if min number of BS = max number of BS (say, 1), how could gain/loss happen? And how StepTime (Fig. 2, Table 5) is defined? What is the purpose?
- Replacement turnover: if min and max number of sites are not equal to 1, then the replacement turnover is not well-defined. Ex, min = 1, max = 2, and the ancestral has 2 sites, but the descendent has 1, how is replacement defined?
- Functional homology: to measure the alignment performance, contradict the general purpose of alignment.

Criticisms:

- Constraint of individual BS: unless the range of BS abundance is very narrow, individual BS will be relatively unconstained. The behavior of the simulator: the number of BS quickly (under neutral rate) reduces to min and constrained after that.
- The adaptive selection of BS cannot be modeled: all sites will be treated equal. The idea of the emergence of weak sites by neutral random walk followed by selection is missing.

- The parameters of the simulator is generally unknown/hard to estimate, thus the relevance of the simulator to real world is unclear (for example, the range of BS abundance has a hugh effect on the BS turnover rate). BS abundance parameters: no way to estimate; spatial parameters: not even validated in most cases
- Not possible to use for bioinformatic studies because of global dependence. E.g. for CRM prediction

Remark:

- Need to check the simulation procedure: the correct way is to estimate the rates first (in this case, the rates of the mutations that violate the constraints will be zero) then to sample the event. Not sample the event according to the neutral rates and then reject the event if it violates the constraints.
- EMMA-simulator addresses these problems by:
  - Model the evolution of individual BS while at the same time meeting the global constraints through balancing gain and loss rates (spatial constraints are generally weak)
  - The parameters can be automatically estimated from data
  - The model can be directly used to aid bioinformatics studies: CRM prediction, sequence alignment
- One needs to know the fitness function in order to properly simulate the CRM evolution, when a new TFBS occurs, whether it will be accepted (constrained thereafer) will be determined by this fitness function. PSPE: never accept a new site if the number of sites = min number of sites; EMMA: always accept a new site.

9. Affinity threshold model [Lusk & Eisen, PSB, 2008]

   Aim: the evolutionary pattern of TFBSs - dependence on the threshold of binding affinity.

   Methods:

   (a) Affinity threshold model: random mutation on TFBS, any mutation that makes the TFBS below the threshold will be discarded.
   (b) TFBS identification: use ChIP-chip data of 124 yeast TFs (with PWM available), scan the binding regions with the PWM for the strongest hits as the TFBSs, thus each region has only one site.
   (c) Measuring evolutionary rate: the substitution rates by parsimony cost.

   Results:

   (a) Affinity threshold model predicts the relative substition rates in different positions. By using an appropriate threshold, it can fit better than HB model.

10. Yeast promoter evolution [Raijman & Tanay, PLoSCB, 2008]

   Problem: a formal model of promoter evolution incorporating explicitly the strength of selection

   Methods:

   (a) Model (RST model): mutation (neutral rate) occurs at any position in the promoter sequences. If there is no BS gain or loss, neutral (i.e. accept the mutation); if create or a new BS or destroy an existin BS for TF t, then this event is fixed with probability σ_t, called selection factor.
   (b) Likelihood computation: the likelihood of an entire promoter can be decomposed into epistatic blocks. The transition prob. of each block: $P(x \to y|t)$ is done by first find all possible intermediate states z, and compute by dynamic programming on discrete-time MC:

$$P(x \to y|t) = \sum_z P(x \to z|t-1)P(z \to y|\text{in } dt) \tag{3.6}$$

(c) Learning the strength of selection of various TFs: the parameters are esitmated by MLE; greedy approach, each time add one TF, it is accepted if the LL score is increased.

(d) Estimate the strength of selection of a set of k-mers (or a type of BS): measured by the ratio of the observed number of conserved motif appearances and the expected number under neutral model. If this type of BS is under stronger selection, it should have larger ratio

(e) Data: yeast promoters in different species pairs with Scer, motifs (245 from literature + learned from data)

Results:

(a) Strength of selection of various TFs: constant over different divergence times; weak correlation with redundency: the TFs with homotypic clustering tend to have lower selection factors.

(b) Strength of selection of different types of BS: (i)the strong k-mers are clearly under selection; (ii) the boundary k-mers are under weaker selection, but still significant; (iii) the substitution between strong k-mers and boundary k-mers are much less than under neutral expectation

Remark:

(a) Any motif appearance must be a TFBS

(b) Gain and loss should be treated differently (no distinction of positive and negative selection): for example, if a TFBS is lost previously, then a gain of TFBS will actually be adaptive

(c) All TFBS of the same TF in all promoters have the same selection strengh

11. Waiting for two mutations [Durrett & Schmidt, Genetics, 2008]

Problem: what is the time of fixing two compensatory mutations (i.e. the first one is deleterious, which is compensated by the second mutation)?

Model:

(a) A sequence with two sites: a or a' and b or b', where $'$ indicates inactive site. We are interested in this pathway: $ab' \to a'b' \to a'b$, where ab' is the w.t.. Let u_1, u_2 be the rate of the first and second mutation (call A and B mutant), respectively.

Results:

(a) Drosophila: $N = 2.5 \cdot 10^6$, $u_1 = 10^{-7}$ (BS loss), $u_2 = 1/3 \cdot 10^{-8}$, (i) if A mutation is neutral and B mutantation is mildly advantageous, $s = 10^{-4}$, then time is $400,000$ years; (ii) if A mutation have fitness $r < 1$, then the waiting time is increased, e.g. $r = 10-4$, then increase is about factor 2.

(b) Humans: $N = 10^4$, require more than 100 million years.

12. 1001 ways of making similar enhancers [Veitia, BioEssays, 2008]

Hypothesis: compensatory mutations can create many enhancers with similar function.

Model:

- Compensatory mutations: a mutation that leads to slightly lower affinity in one site can be compensated by a higher affinity in another site, and vice versa. This process may eventually lead to birth of new sites and death of existing ones. Note that: at every step, the output of the enhancer is constant (neutral or nearly neutral).
- The effect of recombinations: the enhancers may be polymorphic, thus recombination may lead to new enhancers. However, those that significantly change the length of enhancers tend to be purified.

13. Evolution of spatial patterning [Khatri & Sear, PNAS, 2009]

    Problem: the ability of a regulatory system to adapt to optimal phenotype? And how it is affected by mutational entropy (i.e. if there are too many mutations, the system may not stay at the good phenotype).

    Background: in steady-state evolutionary processes, there is a balance between mutational entropy S and mean fitness $\langle F \rangle$, where a quantity analogous to the Helmholtz free energy, the free fitness $\Phi = \langle F \rangle + \frac{1}{\nu} S$ is at maximum, where ν depends on population size.

    Model:

    (a) Spatial patterning: input morphogen gradient $M(x, \alpha)$, which is an exponential function with the paramter α (steepness); and the output 1D gradient, $T(x)$ is the phenotype.
    (b) Genotype-phenotype map: the genome contains binding sites of RNAP and of morphogen. The morphogen binds to its site, then through interaction with RNAP to stimulate transcription. The output is determined from promoter occupancy by RNAP [Shea & Ackers]. Only the simple case of two sites, binary genome.
    (c) Fitness function: the target pattern is high expression in anterior, but low in posterior. Thus the fitness is proportional to the total amount of anterior expression minus the total amount of posterior expression (with normalization).
    (d) Evoluionatry simulation: both mutation of the genome, and the "continuous" mutation of the morphogen gradient α (chosen from a Gaussian ditribution). Assume the population is monomorphic.

    Results: overall, always converge to an equlibrium that consists of an ensemble of significantly fit solutions. However, substantial complexity:

    (a) At very small population size ($N = 20$): converge to suboptimal phenotype, corresponding to $\alpha \approx 7$. In free fitness landscape, single peak at $\alpha \approx 7$.
    (b) At intermediate population size ($N = 110$): one suboptimal phenotype ($\alpha \approx 7$) and optimal phenotype ($\alpha \approx 10$), the probability of converging to the optimal phenotype is greater than 0.9. In free fitness landscape, two peaks at α equal to 7 or 10.
    (c) At large population size ($N = 500$): each independent run is trapped to a different local optimum.

    Question: the role of trans- level mutations (i.e. mutation of α)?

    Remark:

    (a) Evolutionary dynamics of monomorphic populations: could be viewed as random walk in the sequence space, thus follow the usual Markov chain formalism. Then the questions can be posed as, e.g. the probability of converging to the optimum phenotpye, etc.
    (b) The limitations of this study:
        - Population size and monomorphic assumption: not held in general, as the length of promoter/enhancer is not considered.
        - The relevance of results: in the regime of $N < 500$, much smaller than real populations.
        - Oversimplifications of the system: only single pattern (switch-like); typically involve interactions of multiple TFs (activators and repressors); etc.

14. Evolutionary mirages [Lusk & Eisen, PG, 2010]

    Hypothesis: the enrichment and conservation of overlapping and clustered binding sites are results of evolutionary process, not direct selection on these structure. Namely, mutations (esp. deletions) that act on overlapping or adjacent sites tend to have a larger consequence, and eliminated often.

    Methods:

(a) Simulation: (1) single sequence: each TF must have a minimum number of sites, mutation-selection cycle, where deletion bias is introduced. (2) population simulation: $N = 10,000$, fitness defined as $1 - ks$, where k is the number of sites below the threshold, and s is a penalty term. Only single sequence results were reported, as the population simulation lead to similar results.

Results:

(a) BS turnover rates: depend on the specificity, length and GC content of the motifs.

(b) Overlapping sites: longer half life (about twice) than isolated sites.

(c) Clustered binding sites: with deletion bias, (1) enriched in the simulation results; (2) more conserved than isolated sites. However, without deletion bias, the effect of conserved clustering is minimal as the frequency of multi-site deletion is low.

15. Redundancy and multiplicity of enhancers [Paixao & Azevedo, PLCB, 2010]

Goal: explain the commonly observed multiplicity in enhancers. Examine how different factors enhancer multiplicity, including: the selection of redundancy (e.g. that make expression more robust), TF binding degeneracy, recombination, etc.

Background: eve stripe 2 enhancer has 5 Bcd sites, removing any of which changes the endogeneous expression pattern. Orthologous enhancers lack the site Bcd-5 but still drive the natural expression. Thus multiplicity may not be equal to redunancy, but redundant sequences may still be important transitional form.

Methods:

(a) Enhancer model: consider only one activating TF, let $f(m_i)$ be the measure of binding affinity of site i, where m_i is the number of mismatches. And the expression output is simply $F = \sum_i f(m_i)$. Under the simple model, $f(m_i)$ is 0 whenever $m_i > 0$; under the model of TF promiscuity, $f(m_i)$ is a step function allowing mismatch.

(b) Selection scheme: Full redundancy model - the fitness of having two sites is equal to one site; partial redundancy model - the fitness of having two sites is 1, and one site is $1 - s$.

(c) Evolution: for each site, the number of mismatches as the state, thus mutation-selection balance of a system with L states (L is the length of motif, thus maximum number of mismatches). With the number of binding sites, $K = 2$, the 2D system of L^2 states. What is interesting is the fraction of redundancy states (i.e. 2 sites) in the equilibrium distribution.

Results:

(a) Full redundancy, not allowing mismatch: redundant genotype is approximately 2-fold overrepresented (as they are more connected in the neutral network), however the total fraction of redundant states is still very small.

(b) Partial redundancy: selection of partial redundancy (which favors more sites) leads to higher equilibrium frequency of redundant states.

(c) Recombination: a large effect on the frequency of redundant.

(d) TF promiscuity: allowing mismatches, increase frequency of redundant.

(e) Analysis of yeast promoters: define multiplicity of promoters (based on PWM matches), and correlation with various features of promoters/genes. Positive correlation with recombination, promoter length; negative correlation with expression robustness.

Remark: limitations of the work

- "A major challenge for future work is to consider the simultaneous evolution of sites for activators and repressors in the same promoter."

- “The other assumptions are not particularly realistic: synergistic effects among binding sites are commonplace and many TFs are not uniformly promiscuous. The extent to which changing the assumptions of our model would modify our conclusions is not clear at present, and remains a fundamental question for future modeling.”

3.8 Evolution of Regulatory Sequences: Bioinformatics

1. CONREAL: conserved regulatory element anchored alignment [Berezikov & Cuppen, GR, 2004]

   Methods:

   (a) Procedure: (i) PWM hits in both sequences; (ii) compare the homology score (PWM hits + 10bp flanking sequence) of all pairs ⇒ the best ones will be chosen as anchors; (iii) connect anchors to form the alignment

   (b) Data: reference set of known TFBS in human, mouse, rat and fish

   (c) Programs to compare: LAGAN, AVID

   Results:

   (a) Human-mouse-rat comparison: three programs largely agree with each other

   (b) Validation of extra (7) TFBS predicted by CONREAL in human-mouse-rat comparison: the extra TFBS is also conserved in fish (or other outgroup)

   (c) Mammal-fish comparison: CONREAL predicts more TFBS than LAGAN and AVID. Many of these extra TFBS are also found in other mammals; specific case: Foxa2 promoter, the known CREs are aligned by CONREAL but not others

   Remark: a strategy for assessing TFBS prediction is: if the predicted TFBS is true, then it is likely to be conserved in other species. Thus using conservation (additional conservation in this case) to verify a TFBS. This is a general idea of assessing evolutionary inference, add additional species.

2. StubbMS [Sinha, BMC Bioinformatics, 2004]

   Problem: genome-wide scan for putative modules for a given set of TFs

   Methods:

   (a) Window processing: a window (to be scored) is divided into aligned blocks (with high percent identity) and unaligned regions

   (b) Scoring: LRT where aligned blocks are scored as single units using evolutionary models while unaligned regions are scored as if from single species

   (c) Evaluation: the putative modules should be close to genes which have AP expression patterns. Data: 2167 genes from BDGP, 286 AP-patterned genes are positive and the rest negative; procedure: map each predicted CRM to a gene (nearest gene, < 20 kb), and then test the classification of genes

   Results/Evaluation:

   (a) StubbMS performes significantly better than StubbSS: typical improvement over 20%. To predict 100 (+) genes, StubbMS needs 267 predictions while StubbSS needs 343; use the cutoff of StubbMS = 10, predict about half of the 286 (+) genes

   (b) Effect of tandom repeats: masking tandem repeats slightly improves the performance of both StubbSS and StubbMS. Probably due to the widespread presence of polyA/T, which resembles the Hunchback and Caudal PWMs.

3. eCis-Analyst [Berman & Eisen, GB, 2004]

   Probelm: predict CRM or classify true CRMs from negative ones (found experimentally)

   Methods:

   (a) Data: predicted (putative) CRMs using TFBS clustering ($\geq$ 13 sites/700bp). Classified as (+) and (-) sequences according to whether it could drive reporter gene expression. 15 (+) and 18 (-) sequences
   (b) Conservation of TFBS clustering: defined as density of aligned or preserved BS in Dmel and Dpse alignment (LAGAN). Aligned TFBS: overlapping; preserved TFBS: not aligned, but there exists a site of the same TF in the orthologous sequence
   (c) CRM classification: use TFBS density as the score, number of sites/kb. Performance measure: at different score cut off, the number of TPs and number of FPs.
   (d) Genome-wide search of putative CRM (pCRM): rank by TFBS density normalized by the score of random non-coding sequences $\rightarrow$ z-score

   Results:

   (a) All (+) sequences are $<$ 20kb from TSS, all (-) sequences except one are ¡20kb away
   (b) Conservation of CRM in Dmel and Dpse using PID: (+) sequences are more conserved than (-) ones (significant), but insufficient for discrimination
   (c) Conservation of TFBS clustering discrminates (+) and (-) sequences: density of aligned TFBS: 13.8 sites/kb vs 6.8 sites/kb; density of preserved TFBS: 4.3 sites/kb vs 2.2 sites/kb

   Remark: CRM architecture vs composition in detecting CRM: composition alone (plus conservation) seems to be good enough. However, the dataset may be biased (ascertainment bias): all sequences in this dataset have a large number of TFBS. If the TFBS density is low, then the aritecture may be important.

4. Simulating non-coding sequence evolution [Pollard & Eisen, BMC Bioinfo, 2004]

   Aim: create a benchmark tool based on simulation for testing alignment programs.

   Methods:

   - Simulation procedure: modifier version of ROSE, using different rates at different regions with HKY model - neutral sequences and constrained blocks. Sample ancetral sequence from estimates of Dmel non-coding sequences (7-th order Markov chain), then evolve in two equal-length branches.
   - Simulation parameters: divergence from 0.25 to 5.0 (2.24 divergence between dmel and dpse); substitution : indel = 10 : 1; sequence length = 10K; and constrained blocks average size 18bp with density= 0.2

5. Assessment of alignment accuracy and TFBS conservation via CisEvolver [Pollard & Eisen, BMC Bioinfo, 2006]

   Problem: what affects the multiple alignment accuracy and its implications to the inference of TFBS conservation and divergence estimation.

   Methods:

   - Ancestral sequence: either genomic background (sampled) or known CRM sequences
   - Background sequence simulation: HKY85 for substitution, indel events following the Poisson event model and the empirical indel length frequency distribution.
   - TFBS simulation: HB model, no indel events allowed; and any indels from the background that extend to TFBS will not be allowed

- Annotation of TFBS: PASTER with p-value cutoff = 0.001. Used for assessing TFBS conservation

Results:

- Alignment accuracy:
  - Presence of binding site increases alignment accuracy (bell shaped: more improvement at moderate evolutionary distance)
  - Accuracy of multiple alignment is determined almost exclusively by the pairwise divergence of the two most diverged species and additional species have a negligible influence on alignment accuracy
  - Accuracy varies across branches in a tree: most accurate for alignments of sister taxa and least accurate between internal nodes that align sub-alignments
- BS alignment accuracy (defined as the portion of binding sites that are correctly aligned):
  - Misaligned even at short evolutionary distance
  - Better than overall alignment (strong correlation between the two)
  - The two scores (perfect and overlapping) are very different =¿ BS may not be strong anchors
  - BS density, length increase BS alignment accuracy

Discussion: D. mel vs D. pse (div = 1.79): only about 40% of truly conserved BS will be correctly aligned (overlapping)! Therefore, unless TFBS turnover rate is much higher, the turnover rate estimation based on the fixed alignment is almost completely misleading.

Remark:

- Large space of improvement in alignment of sequences with binding sites
- To analyze a problem of evolutionary inference, separate the effect of alignment and the inference methods. Compare the methods by using the true alignment from the simulator.

6. Comparative promoter analysis in cell junction-associated proteins [Cohen, PNAS, 2006]

   Methods:

   (a) Promoter framework construction: seed gene → orthologous sequences in multiple species → promoter framework (FRAMEMARKER) using a TFBS motif library -¿ check if the framework matches additional seed genes.

   (b) Scan for novel target genes: scan for the promoter framework in the whole genome and check conservation

   Results:

   (a) NPHS-1 and ZO-1 share the same promoter framework

   (b) In 6 genes, the framework is conserved in at least 2 species

7. CisPlusFinder [Pierstorff & Wiehe, Bioinfo, 2006]

   Methods:

   (a) Sstrategy: a sequence that contains a local clustering of conserved blocks, as well as a local overrepresentation of words will be candidate of CRM. (i) PLUS (perfect local ungapped sequences); (ii) select PLUS: the minimum length of PLUS (from random sequences); and there must be a core motif of this PLUS that is locally overrepresented; (iii) adjacent PLUS's are merged to give the score of a sequence.

   (b) Parameter training: using a small CRM dataset from Paptsenko

   (c) Data: use Dmel + Dyak, Dana, Dpse and Dvir

(d) Evaluation: CRM classification using HexDiff & REDFly dataset. The metrics are taken from HexDiff

Results:

(a) In HexDiff, CisPlusFinder has a higher sensitivity than all other methods, but low sensitiviy. Stubb using Dpse or Dvir performs well, high sensitivity and better specificity than CisPlusFinder. Overall Stubb is similar to eCis-Analyst

(b) Correlation of PLUS with known TFBS (from FlyReg) is low

(c) Analysis of false negatives from CisPlusFinder: require perfect alignments: which may not be available for divergent species; CRM may be in the region of high substitution rate

8. SimAnn: simultaneous alignment and annotation [Bais & Vingron, Bioinfomatics, 2006]

Methods:

(a) Algorithm: modified local alignment. The aligned binding sites are scored using log-likelihood ratio of binding site evolution vs background evolution. Substract a "profile penalty" (roughly significance cutoff of the LRT)

(b) Baseline methods: ex. ConSite, CisOrtho alignment and extraction of conserved regions; scan for TFBS in the conserved regions.

(c) Simulation: independent samples of the two TFBS and inserted randomly into the simulated sequence

(d) Data: D. melangoster (799) and D. pseudoobscura (1028), 17 verified BS in Dmel

Results:

(a) Case study: eve strip 2 module. ConSite - 5; multi-step procedure (pre-local alignment + scanning) - 10; SimAnn - 9. Example: Kruppel 4 site - the alignment of UCSC clearly differs considerably from other methods such as ConSite and local alignment.

Criticism:

(a) Considers only independent sample from the PWM, not the evolution of BS

(b) Simulation data: the substitution model is unrealistic, not know the indel model

9. EEL [Hallikas & Ukkonen, Cell, 2006]

Problem: same as StubbMS. Pairwise comparison

Observations: (Figure S1 C and D)

(a) TFBS clustering in mammals is not as strong as in Drosophila (especially homotypic clustering) → reduce the signal/noise ratio

(b) The much larger size of mammalian genome further reduces the signal/noise ratio

Methods: score a region by affinity of putative TFBS in both species; and score clustering of TFBS in the two sequences

(a) Scoring: affinity score = $\lambda(W_i + W_j')$, where W_i, W_j' are affinities of the two TFBS. Distance penalization: let x, x' be the distance of the two TFBS, then $F(x, x')$ penalizes if x and x' are large; if x and x' are very different; if the angles of the TFBS are different. Unaligned TFBS will not contribute to the score.

(b) Local alignment algorithm: first scan for putative TFBS in two sequences; then treat each TFBS as a single unit and apply the DP alignment algroithm to maximize the score; each time one best local alignment is found, then recursively apply DP

Note: since distance penalization only depends on adjacent TFBS, then DP is applicable.

Results:

(a) Data preparation: orthologous genes → extract flanking seq's; scan for TFBS

(b) Drosophila (Dm vs. Dp): EEL on eve flanking region, found all 4 known eve CRM's

(c) Human-mouse: genome-wide scan on 20K genes for the TFs: GLI1-3, Tcf4 and c-Ets1 ⇒ predicted CRM.

(d) Validatation of predicted CRM: could drive the expression of some reporter gene, and moreover, the expression pattern of the reporter gene is similar to that of the target gene of this CRM

Remark: because the spacing sequences are not aligned, thus the distance penalization will be essential: emulate the effect of aligning spacing sequences. In other words, if the spacers in the two sequences have the same or similar length (i.e. better aligned), then it will have less penalization.

10. TF-map alignment [Blanco & Guigo, PLoSCB, 2006]

Problem: find the genes with promoter similar to some given gene (coexpressed); or find the transcriptional regulatory "superpattern" or architecutre from a set of coexpressed genes

Methods: construct TF map by PWM scanning, then align the TF map using the restriction enzyme map alignment

Results: evaluation: given a gene, assess whether its coexpressed gene can be found through promoter comparison

Remark: different from alignment of orthologous CRMs because there is no conservation, rather, alignment depends on the existence of the consensue architecture.

11. eSimAnn [Bais & Vingron, J Biosci, 2007]

Methods:

(a) Algorithm: improve SimAnn by defining the score for TFBS using the HB model (LRT)

(b) Evaluation data: human-mouse orthologus sequences with 98 verified TFBS

Results: comparison with MONKEY: similar performance. At lower p values, eSimAnn is slightly better

12. EDGI [Sosinsky & Califano, PNAS, 2007]

Problem: given a sequence and its orthologs, test if it is a CRM without using motif knowledge

Methods: find putative motifs through homotypic clustering, then check if they are conserved

(a) SCM (short conserved motifs) discovery: by SPLASH, a word is SCM if it occurs in at least q species with density at least k and length l

(b) Find the maximum window of size L in all sequences which contain a set of common motifs, possibly in different order, spacing, number of copies.

13. MEC and BLS [Stark & Kellis, Nature, 2007; Kheradpour & Kellis, GR, 2007]

Aim: discovery new motifs and prediction of individual TFBSs through inter-genome comparison.

Methods:

(a) Sequence alignment: two programs MAVID and Blastz/Multiz.

(b) BLS: branch length score defined on a motif instance. As percentage of the entire phylogenetic tree of 12 species.

(c) Confidence value of BLS: map the BLS of a motif to a confidence value. Suppose at a BLS level, e.g. 0.5, we predict P instances in certain regions (e.g. promoters), and predict average N instances of control motifs (random, the same frequency as the motif). Then the confidence level is $1 - N/P$ (N among P instances are false positives). This is simply a conservative estimate of FDR. Note that the BLS cutoff for a given confidence level is motif-specific. For Snail, 60% confidence - BLS cutoff about 0.1.

(d) Motif discovery: assess the excessive conservation (MEC) of motif instances relative to overall conservation in promoters, intergeneic regions, etc. Note that motifs are searched in a specific type of regions so that the broad functional role can be inferred (e.g. if enriched in promoters, then likely to be invovled in transcriptional regulation).

(e) Functional test of motifs: test if a motif is enriched or depleted in a certain set of genes. For each gene, whether it is a target of a motif is determined by: a motif instance at the promoter (2000bp) or intron with 50% BLS. Then the enrichment is tested by the hypergeometric test.

(f) ChIP-chip data: Met2, Twist, Snail, and direct targets of CrebA

(g) Target gene prediction (for constructing regulatory networks): a motif instance in the promoters or 5' UTR for TF motifs, and 3' UTR for miRNA motifs. Cutoff: BLS confidence level 60%.

Results:

(a) Motif discovery: 145 pre-transcription motifs, recovering 40 (46%) of the 87 known transcription factors. The AP motifs found are: bcd, cad and gt. In fact, Among 74% of AP motifs did not exceed the conservation expected by chance in promoter regions.

(b) Functional preperties of predicted motifs: 75 (52%) were either enriched or depleted in genes expressed in at least one tissue (ImaGO), compared to 59% of known motifs and 3% of random controls. Also significant enrichment in groups of genetic interactions (FlyBase), KEGG or GO. Overall, 68% of discovered and 70% of known motifs were enriched or depleted in one of the functional categories.

(c) Positional constraint of motifs: 15 of the discovered motifs (10%) were significantly enriched near transcription start sites (compared to 14% of known and 1% of random motifs). Several were enriched at precise positions and preferred orientations

(d) Predicted motif instances tend to occur in promoters instead of other regions.

(e) Validation of sites predicted by BLS: (i) enrichment of known in vivo targets in conserved motif instances increased sharply for increasing confidence values (high precision) - Fig. 7b; (ii) a large fraction of motif instances in experimentally determined target regions was conserved: e.g. 76% of motif instances in direct CrebA targets were recovered (high sensitivity) - Fig. 7c.

(f) Comparison of ChIP (alone) and conservation for TFBS prediction: comparable level of enrichment of TFBSs (non-conserved motif instances in ChIP bound regions show only 1-2 fold enrichment). The sites that are both conserved and bound show strongest enrichment (Fig. 7d).

(g) Effect of multiple alignment on prediction of motif instances: (i) at 60% BLS confidence, the agreement of two programs find 59% TFBSs vs 47% if require perfect conservation; (ii) Blastz/Multiz alignment identified 11% more TFBSs (unlikely due to lower specificity of Blastz/Multz as the confidence measure has corrected for this).

(h) Effect of motif movement: when defining BLS, allow for motif movement in orthologous sequences. At the same confidence level, change the motif movement parameter and observe the number of motif instances (sensitivity). Motif-specific optimal movement, e.g. for Bcd, about 400 - 500 bp. A single best value for all TF motifs: 20bp.

(i) Choice of species for TFBS prediction: BLS with all 12 species vs. additional requirements of Dmel-Dpse conservation, full conservation in melanogaster subgroup (4 species), or full conservation in Sophophora subgroup (9 species). At the same confidence level, BLS is the most sensitive (about 1.5 fold more instances are discovered).

(j) TRN in fly: at 60% confidence level, containing 46,525 regulatory connections between 67 transcription factors and 8,287 genes. Genes with the highest sites are those involved in development (all stages), and lowest are ubiquitously expressed genes and maternal genes with housekeeping functions.

Remark:

(a) The species choice: only comparison with methods with full conservation. If allowing lineage-specific changes, 12 species always better than, say 9 species?

14. EvoPromoter [Wong & Nielsen, Bioinfo, 2007]

Methods:

(a) Model: CRM model: HMM that emit background or TFBS; background substitution: HKY85; TFBS evolution: HB model where the mutation rate matrix is equal to the background substitution matrix.

(b) Parameter estimation: CRM parameters (HMM transistion probabilities): by Baum-Welch algorithm; evolutionary parameters: trained from PAML (branch lengths and transition/transversion bias)

(c) Procedure: alignment: CHAOS+DIALIGN; window scoring: scan the sequence with a window of size W, find the Viterbi path (annotation) and the number of predicted TFBS in the window; cutoff: find mean and SD of the window scores, and the cutoff = mean + 1 * SD

(d) Simulated data: sample 10kb sequences and evolve according to the background evolution model. Plant randomly 3 CRMs of size 500bp inside and evolve according to the CRM evolution model (CisEvolver), with 9 TFBS per CRM. The tree: from Pollard tree

(e) Real data: 16 genes with upstream 10kb each (from [Chan & Kibler, BMC Bioinfo, 2005]). CRMs from [Schroeder, PLoS Biology, 2004] are (+) data, but no (-) data. Use the same 9 PWMs used by [Schroeder, PLoS Biology, 2004]. 5 Drosophila species, including distant ones like Dpse and Dvir

(f) Performance metrics: CRM-level prediction: the number of correctly predicted CRM (sensitivity) and positive predictive value (PPV), defined as TP/(TP+FP). PPV can be interpreted as the probability that a predicted CRM is a true one. Nt-level prediction: also sens. and PPV, defined over nt. columns.

Results:

(a) Simulated data: (i) predict 28/30 CRMs if using true alignment, compared with 17/30 by MCAST and 23/30 by MSCAN; (ii) predict only 20/30 if doing the alignment by CHAOS+DIALIGN

(b) Real data: (i) CRM prediction: EvoPromoter has a slightly higher sensitivity, but overall similar; (ii) nt-level prediction: EvoPromoter performs slightly worse

Discussion/Criticism:

(a) The performance of EvoPromoter strongly depends on the alignment from the simulation

(b) The species chosen may not be the best: the species are too distant s.t. either alignment has poor quality or/and the CRM conservation is too low to help prediction

(c) Simulation: the branch length is based on the neutral substitution rate while the actual rates in CRM or neighboring sequences are much lower

15. Predicting regulatory target genes of TFs in fly [Aerts & Hassan, PLoS ONE, 2007]

Goal: given a TF, predict if some gene is the target of this TF or not via its CRS

Problem: evaluate in particular 4 issues about target gene (TG) prediction

- Do TFBS really form homotypic clusters, to what extent?
- All methods rely on PWM, can one improve PWM by utilizing orthologous sites?
- Can one improve TG prediction by utilizing the co-occurrence of motifs of interacting TF(s)?
- What is the best way of using conservation information: sequence (nt.) level conservation or network-level conservation?

Methods:

(a) Data: 34 TFs of Dmel and 166 target genes, extracted from known TFBS in FlyReg

(b) Evalution strategy:

  - Running the methods: because all methods need training, so run the methods with cross-validation (LOOCV). In particular, first train the enhancer model from all other TG and then score the left-out target. Because the scores are not normalized, use random negative sequences, and the rank ratio of this score among all negative sequences as the final score (empirical significance: 1-spec)
  - Performance metric: detected fraction (sens) rank ratio (1-spec) ROC curve. And use AUC if a single number is desired

(c) phyloPWM: add orthologous sites to the existing ones to construct the PWM (no evolutionary modeling)

(d) Score putative CRM by conservation: (i) sequence level conservation: functional sites should be more conserved than background. In particular, use phastCons scores with certain threshold or Stubb. (ii) network level conservation: a functional CRM may be conserved, i.e. present in all ortholgous sequences in all species, but not necessarily at nt. level. Compute the score of each orthologous CRM, and then integrate all scores to form a single one using order statistics (combine species-specific ranks)

(e) Heterotypic clustering: learn the co-occurred motif from the sequence of all training TG by a motif sampler; add the new motif to the enhancer model

Results:

(a) Baseline method: score putative CRM with trained enhancer model using known PWM (Cluster-Buster). Could reach a good performance when PWM is constructed from known sites

(b) phyloPWM on average does not improve the TG prediction. But it improves significantly 5 individual TFs. These are TFs with a small number of known sites

(c) Network-level (NLC) vs sequence-level conservation

  - Stubb-MS on Dmel and Dpse is significantly better than Cluster-Buster on Dmel alone
  - Cluster-Buster on Dmel and Dpse better than Stubb-MS at rank ratio < 0.2, but worse after that
  - Cluster-Buster NLC on all 11 species significantly better than all pair-wise NLC
  - Masking all nts in Dmel with phastCons score < 0.9 (or 0.7, 0.5) over 11 species $\Rightarrow$ better than Stubb-MS and NLC on 2 species, but slightly worse than NLC on 11 species

(d) Use extra motifs/interactions between motifs:

  - Motif prediction from training TG and orthologous sequences of 11 species with PhyloGibbs and oligo-analysis as the enhancer model: not as good as using known PWM
  - Adding NLC with the above methods improve the performance
  - Adding true PWM with NLC and PhyloGibbs or oligo-analysis is best: better than NLC + motif; and better than know PWM + NLC

Discussion/Remark:

(a) Homotypic clustering is not studied in a controled fashion: only know PWM is important, but not know how important the homotypic clustering is.

(b) phyloPWM method could be improved, not considering the evolution of TFBS

(c) Stubb-MS is not as good as NLC on the same 2 species: is the difference significant? Is fixed alignment of Stubb-MS a problem?

(d) PhyloGibbs and oligo-analysis fail to discover the true PWM in most TFs (phyloGibbs: 8/34; oligo-analysis: 14/34), possible to improve?

(e) The power of extra motif need more examination: want to compare known PWM vs known PWM + extra motif(s).

16. CSMET [Ray & Xing, PLoS CB, 2008]

Aim: detect conserved, including partially-conserved, TFBSs within CRMs.

Methods:

(a) TFBS tunover model: the functional state of a block evolves according to a functionality substitution model (JC model). The block probability under a given assignment of functional states for all leaf nodes is factorized into two parts: (i) the marginal tree that consists of only BSs: T_m; (ii) the marginal tree that consists of only background: T_b. Summation over all 2^M configurations of functional states of leaf nodes (M is the number of species).

(b) CRM model: 3-state HMM - one background nt., two states for motif (both strands).

(c) Posterior inference of BS positions: for any column, computes its probability of being a start position of some binding site blocks. The block probability is computed by the above procedure, while the rest by standard forward-backward algorithm.

(d) ML training: annotated alignment (positions of all binding sites), separate estimation of nt. substitution parameters (both backgrounds and motifs), and functionality substitution parameters. Evaluation is done in a K-1 cross validation scheme (K-1 used for training, 1 for testing).

(e) Sequence simulator: to generate a motif block, (i) simulate a sequence block under a complete motif tree; (ii) simulate a sequence block under a complete background tree; (iii) simulate the functional tree (1s and 0s); (iv) merge the complete motif block and background block according to the functional tree: i.e. choose the sequence from the motif block if it is 1 under the functional tree, and from the background block if 0.

(f) Data: 14 fly CRMs, 250 binding sites. Multiple alignment is from UCSC. 11 species without wil.

Results:

(a) Comparison with other programs using simulator, and PSPE simulation (no BS turnover, all lateral displacements are removed).

(b) Comparion on real CRM dataset and prediction of partially conserved TFBSs: CSMET precision is generally lower than other programs such as PhyloGibbs and Stubb (0.1-0.3 for bcd, hb, Kr, kni and < 0.4 for cad), but has much higher recall (> 0.5 for all and > 0.9 for bcd). Overall, it has higher F1 than all other methods in all except Kr.

Remark: CSMET significantly over-predicts TFBSs as manifested by its very low precision. Thus the "better" performance is really a different spec-sens tradeoff scheme adopted by CSMET.

Criticisms:

(a) Modeling BS turnover: in fact two separate processes: loss of existing TFBSs and gain from background nucleotides. Not explicitly modeled in CSMET model, two consequences: (i) JC model is clearly wrong, because there is no reason to believe that gain and loss rates are equal and at equilibrium there are equal number of functional and non-functional states; (ii) the gain and loss rates are related by the equilibrium density of TFBSs (the parameters in HMM).

(b) Probability computation of a block: marginalization of two trees is not the correct way of computing the probability because two marginal trees are not independent. In fact, they will overlap, or some part of the tree is missing.

(c) Training: depends on annotated alignment of CRMs, generally unavailable.

(d) Treating alignment gaps: not allow insertions within TFBS blocks. Insertion within a TFBS of any species will make the block undetectable. Greatly reduce the sensitive with a large number of species.

(e) Simulator: the simulation of TFBS block with possible turnover is wrong. Cannot merge the two blocks from evolving each of them separately. Also no indels are supported.

(f) Multiple motifs: not supported.

17. Assessing phylogenetic motif model (PMM) [Hawkins & Bailey, Bioinfo, 2009]

    Goal: use PMM to predict TFBS, and compare the results with PWM scanning.

    Methods:

    (a) PMM methods: PMM score defined as the LRT of HB vs HKY model. The difference among methods: Motiph - ignore regions with gaps; MONKEY and rMONKEY - local realignment, use the adjacent columns to realign the sites.

    (b) Evaluation data: yeast promoters, (1) known sites of 21 TFs from SCPD; (2) random shuffled version of motifs. Assess FDR in both cases at a certain sensitive level (top 20 or top 50 predictions).

    (c) Theoretical accuracy of PMM scanning methods: sample the distribution of the LRT score under theoretical models and calculate FPs and FNs, according to the overlap of distributions.

    Results:

    (a) Known yeast TFBSs: PWM better than PMM. Probably due to the missing of many weak but functional sites.

    (b) Random shuffling: PMM much better than PWM, as expected. Among three programs with different realignment strategy, MONKEY performs the best.

18. PRIORITY: alignment-free method [Gordan & Hartemink, NAR, 2010]

    Problem: motif finding in reference species. The orthologous promoters can be used.

    Methods:

    (a) Framework: Gibbs sampling for motif finding, the conservation information is used for specifying the prior of each position: more conserved positions will receive higher prior probabilities.

    (b) Alignment-free conservation score: to computer the conservation score at any position, first find the K-mer starting at this position; then count the number of orthologous promoters that contain this K-mer (regardless of position and orientation). The fraction is the conservation score.

    (c) Alignment-dependent conservation score: either PHASTCON score, or the similarly defined conservation score where the K-mer count only applies to the aligned positions.

    Results: evaluate the number of recovered motifs in [Harbison04] data sets.

19. Detecting selection in regulatory sequences

    Sunflower [Hoffman & Birney, GR, 2010]:

    (a) Goal: a measure of promoter change that reflects the change of TF occupancy, then use this measure to compare promoters to understand which genes are under stronger/weaker regulatory selection.

(b) Methods:

- Measure of promoter selection - transcriptional divergence: HMM to represent a promoter sequence and the posterior probabilities of motifs thus represent the occupancy profile of this promoter. For a given mutation, the profile will be changed, and the relative entropy of the two profiles is defined as the binding shift of this mutation, t. The average t over all positions of a promoter, T, can be used to define transcriptional distance d_T, which can then be normalized with d_s, $\psi = d_T/d_s$.
- Data: human promoters, comparison with dog genome, all JASPAR core motifs are used in the model.

(c) Results: different functional classes of genes show different patterns of ψ and ω (measure of protein divergence). First, the two measures are not correlated. Second, the class with low ψ and low ω include: sensory organ development, TF activity, etc.

Chapter 4

Adaptation and Evolution of Novel Phenotypes

The basics of adaptation: the evolutionary process that creates novel phenotype for the organism, making it better suited than before. It could be driven by:

- The change of the environment: making novel phenotype more desirable. This include changes of habitat (climate, geographical environment, or life style), that demand new phenotypes; or at the molecular level, changes of the associated genes, cells or tissues, etc., that demand changes of expression pattern or function of genes.
- The internal adapation: novel features evolve spontaneously from random mutations that confer advantages on organisms. Ex. the change of brain that gives language ability; the evolution of immune systems; etc.
- The two scenarios of adaptation are closely related. In the real process, internal adaptation may enable organisms to live in a different habitat (e.g. the change of the limbs to wings may allow the orgnaims to fly), and then the new habitat drives adaptations of all the physiology and developmental plans of organisms (e.g. the flying life style demands change of bones, respiratory system, skin, etc.).

Spandrel and challenge of adaptationism:

- Spandrel: a phenotypic characteristic that is a byproduct of the evolution of some other character, rather than a direct product of adaptive selection.
- Example: Hsp90 and/or other mechanisms that may increase variations of a population. This may serve to increase the chance that some individual will survive some form of selection; however, the mechanism itself is evolved for other reasons (Hsp90: chaperon to fold proteins in the presence of heat shock).
- Example: evolutionary psychology of human behavior such as different sexual behavior of men and women. Women tend to have less sex may be simply a result of minimizing the chance of catching infections.

How do we determine the adaptive values of a new phenotype? Whether a trait is adaptive and what drives the adaptation?

- Example: [Luis Barriero talk] evolution of pygmy phenotype. Hypothesis of its adapative values include: food limitation, thermoregulation (humid), better mobility.
- Signs of positive selection: would support the hypothesis that adaptation drives the change of phenotypes.

- Functions of selected genes: would indicate the likely biologial processes under selection, thus the likely biological hypothesis/factors. Also if we know that some genes are related to a hypothesis (e.g. thermoregulation), then the lack of sign of selection in the gene is evidence againest that hypothesis.

Examples of evolution of complex systems/phenotypes:

- Evolution of photosynthesis [10 Inventions]: (1) two photosystems were formed from gene duplication (similar function, extract electrons); (2) photosystem was evolved from the system for oxidizing hydrogen sulfide; (3) electron transport chain existed before oxidization.
- Evolution of eukaryotic cells [10 Inventions]: (1) reunion of two prokaryotic cells that are symbiotic; (2) new entity evolved ability of phagocytosis (because of higher energy efficiency, etc.), and then features of eukaryotic cells.
- Human naked skin [SciAm, Jan/Feb, 2010]
- Bacterial flagella [E. coli, Microcosm]
- Darwin's finches: the beak size changes with climate change every year.
- Diversity of cichilds in Lake Victoria.
- Evolution of land vertebrates from fish: lobe-fin fishes evovles legs for underwater walking (near the costal regions, where the food underwater may be rich), then move to the land. This illustrates: (1) evolvability of the developmental system - could evolve legs from fins; (2) exadaptation - the legs for underwater walking later used for land; (3) selection pressure, probably environment changes, such as droughts that favors land moving animals.
- Mimicry: some butter flies have wing patterns remarkably similar to other unpalatable species. For the unpalatable species, their distastefulness originates from the obnoxious compounds from the plants they eat (and stored in some gland). The palatable species did not acquire this capacity, but by mimicing the wing patterns of unpalatable species, they gain evolutionary advantage by avoiding predators. Presumably the advantage is greater than the cost of losing the old pattern, which may be important for mating.

Detecting genetic basis of adaptation [personal notes]

- Comparative genomics: lineage-specific genome changes such as duplication. Ex. in human-mouse comparison, many mouse-specific gene clusters encode proteins with roles in reproduction, immunity and olfaction.
- Population genetics: polymorphism data to infer positive selection. Selective sweep - if a gene is under positive selection, then the rare alleles of neutral linked region may show higher frequencies than expected by neutral, due to selective sweep.
- Gene expression comparison: many brain genes show pattern of positive selection in human. Caveat: change of the composition of a tissue (e.g. the relative porportion of cell types) will be accomponied by altered expression profiles, but these are indirect consequences, not cause of developmental changes.
- Candidate gene approach: use functional/genetic data to locate candidate genes of a certain trait (e.g. from medical genetics), and test selection and function of candidate genes. Ex. FOXP2 identified from mutations, and signature of selection was verified.

The genetic theory of adaptation [Orr, NRG, 2005]

- Problem: the genetic architecture of adaptation: major genes of large phenotypic effects or many genes with small effect each? Do populations evolve quickly at first and move slowly? Will adaptation always lead to best fit phenotypes? etc.

- Micromutationism:
  - Darwin, biometricians (Pearson, etc.), Fisher: a character is underlaid by an infinite number of genes, each having an infinitesimally small effect on the character. Thus adaptation proceeds in very fine-grained steps.
  - Evidence: against the micromutationism from both QTL data (where loci with large effect identified) and microbial experimental evolution.
- Models based on phenotypic evolution:
  - Fisher's geometric model of adaptation: an organism is represented as a set of phenotypic characters, each having an optimal value. With increasing mutational size (large phenotypic effects), the mutations are very unlikely. Thus small mutations are the genetic basis of adaptation.
- Kauffman and NK models: evolution in the fitness landscape on sequence space. There may be global optimum and multiple local optima. The ruggedness of the landscape can be tuned by varying two parameters: N and K.
  - Results from NK models: the number of local optima, the probability that a sequence is a local optimum, the average length of adaptive walks, etc.
  - Limitations: the move in the fitness landscape does not reflect reality: either highest fitness or random (not consider random drift); the evolution always start with a random sequence while real starting sequences often have high fitness.
- Gillespie and mutational landscape:
  - Mutational landscape: evolution under strong selection-weak mutation conditions, a population is fixed at any moment.
  - Application of extreme value theory (EVT): the results does not depend much on the exact distribution of fitness of mutations.
  - Results: beneficial mutations are exponentially distributed, thus adaptation is characterized by a few large jumps in fitness (Pareto Principle: the majority of effect is due to a minority of causes); beneficial mutations become increasingly difficult to come by as adaptation proceeds.

Non-adaptative process of genome/organismal complexity [Lynch, PNAS, 2007; Lynch, NRG, 2007]

- Hypothesis: complexity and related features, modularity and evolvability, in higher organisms, are consequences of non-adapative forces.
- Principle: consider a system with multiple possible states, the probability of transition among states depend on selection, mutational bias and population size.
- Non-adaptative forces and population genetics:
  - Mutation operates as a weak selective force by differentially eliminating alleles with structural features that magnify mutational target sizes.
  - The effect of random drifts is larger in organims with smaller N_e (typically higher organisms): much smaller in vertebrates than in prok. Thus selection is much more efficient in prok. organisms. This may limit the complexity and size of their genomes: even non-functional DNA may carry a mutational cost because of possible deleterious gain-of-function mutations; transposons are almost absent in prok. genomes; etc.
- Complexity from non-adapative forces:

  - Transcriptional networks: the spontaneous process of gain and loss of TFBS changes the connection of TRN. Assuming that the gain or loss are neutral, then the number of sites (connections) may be considerably smaller in simple organisms (with smaller gain rates) than in higher organims. This effect itself explains the complexity in higher organisms.
  - Regulatory sequences: duplication of genes or CREs followed by degeneration creates modular regulatory DNAs.
  - Developmental systems drift: closely related species achieve similar morphological structures by different mechanisms.

- Evolvability:
  - The concept of evolvability is based on group selection (a feature that is good for the species will be selected), little support.
  - Enhanced ability to evolve is not necessarily advantageous.

4.1 Macroevolution and Speciation

Domestication of wine yeast [Querol & Barrio, Int J of Food Microbio, 2003]

- Wine yeast genome: mainly diploid (while lab. strains are both haploid and dipoloid), and even polyploid. Man increase the gene dosage.
- Stress adaptation: in general wine yeast are more adaptive to stress conditions in wine making than in lab. strains. The stress include: oxidative stress, raised temperature and starvation during aerobic fed-batch growth; hyperosmotic stress due to high sugar content in the must; alcohol (highly toxic to yeast metabolism) druing fermentation. A common pattern in adaptation from expression studies is: high level of Gpd1 (glycerol synthesis), and low level of Hsp104.
- Adaptation to industrial environment: in general, genes involved in AA and purine biosynthesis show higher expression level.

Making of Home sapiens [Carroll, Nature, 2003]

- The evolution of behavior traits and associated/underlying physical traits: (in the possible chronological order)
  - Bipedalism: body shape, features involved in locomotion, e.g. pelvis and feet, limb proportions
  - Change of diets, etc.: dentition including presence of a chin (about chewing of food)
  - Tool-making: brain size and skull, elongated thumb and shortened fingers
  - Language and social life: brain size and skull, changes in brain regions such as Broca's area and Wenickes' posterior receptive language area (increased in human vs chimp)
  - Long ontogeny and lifespan: prolonger childhood, delayed sexual maturation (correlated with dental development).
  - Other features: reduced hair cover
- Human capacities are more a product of changes in specialized areas than of neuroanatomical novelties: e.g. brain asymmetry (larger Broaca's area in left hemisphere) found in both human and chimps.

Adaptive evolution of antibiotic resistence [Weinreich & Hartl, Science, 2006]

- Problem: five mutations in a particular β-lactamase increases resistence by a factor of $100,000$. What is the evolutionary trajectory? And is it unique?

- Methods: estimating probabilities of trajectories, assume "strong selection/weak mutation" model, i.e. the time of fixation is much smaller than the time between mutations. Thus each fixation is statistically independent.
- Results:
  - Fitness of every possible allele ($2^5 = 32$).
  - Out of 120 possible paths, 102 are inaccessible because some combinations of mutations have negligible or even negative effects. For the remaining ones, a few trajectories take most of the probability mass.
  - Biochemical and biophysical interpretations of sign epistatis: a mutation often has two effects, one may increase antibiotic resistence and the other may reduce thermodynamic stability and increase aggregation. Thus double mutant may enjoy the incrase of antibiotic resistence without reducing stability.

How did Scer evolve to become a good brewer? [Piskur & Compagno, TIG, 2006]

- Background: the fermantation/respiration switch is affected by the activity of Adh (alcohol dehydrogenase). In Scer, Adh1 is mainly an alcohol producer and Adh2 is mainly an alcohol consumer (convert ethanol to acetaldehyde). Adh1 is constitutively expressed, and Adh2 is expressed only when the interal sugar concentration drops (need to use ethanol as energy source, diauxic shift).
- Evolution of alcohol production and utilization:
  - Ancestral ADH gene is mainly alcohol producer. In Scer lineage. ADH duplication (after WGD) produces Adh1 and Adh2. Scer thus can use a "make-accumulate-consume" strategy: fermentation when glucose is abundant, and switch to ethanol consumpation when glucose is exhausted.
  - This may give Scer an advantage over its competitors: fast ethanol accumulation and ethanol tolerance (generally toxic). And could digest ethanol.
  - In Klac, a poor producer of ethanol (Crabtree-positive), but ADH duplication (4 ADHs) in Klac lineage, and KlAdh4 gene is involved in consuming ethanol from either the medium or previous intracellular production.
  - Possible selective force: about 50-100 Mya (before WGD), abundance of fruits, and thus fermentable sugars, selected fermentation (faster than TCA cyle), whose product, ethanol, has antiseptic effect (inhibit the growth of other microorgranisms).

WGD increases glycolytic flux [Conant & Wolfe, MSB, 2007]

- Background:
  1. Adaptation (gene duplication) related to high-glucose environment.
     - Regulator of respiration: Hap4 (glucose-repressed activator of respiration), before WGD (conserved in Klac).
     - Fermentation pathway: Adh1/Adh2 duplications (after WGD).
     - Glucose repression of other sugars: Rag4 in Klac and Snf3/Rgt2 in Scer (paralogs of Rag4 after WGD). The Rag4 gene occurs before WGD.
     - Pentose phosphate pathway: 4 enzyme duplicates are retained after WGD (this pathway has important role in biosynthesis).
  2. Gene dosage effect: in some cases, increasing the level of one enzyme in a pathway does not lead to the increase of flux; instead, the whole pahway needs to change simultaneously. This provides some rationale for the advantage of WGD.
- Hypothesis 1: WGD followed by gene loss raised glycolytic enzyme concentrations

  - Glycolytic enzymes: 5 (out of 10) duplicates are retained after WGD.
  - Hexose tranporters (major rate-limiting step in glycolysis): 18 copies (many duplicates are retained, plus some post-WGD duplicates).

- Hypothesis 2: The increased enzyme concentrations (glycolytic flux) increase relative flux through fermentation.
  - Respiration is constrained by additional factors: oxygen and mitochondrial (whose genome is not duplicated).
  - The branch point enzymes: PDH for respiration and PDC for fermentation, have different kinetic properties. At low pyruvate level, PDH is favored, but at high pyruvate level, PDC is favored (cooperative activation by substrate, etc.)
- Hypothesis 3: Natural selection (high glucose level) favors inefficient but fast fermentation over efficient but slow respiration.

Experimental evolution of yeast under nutrient limitation [Gresham & Dunham, PG, 2008]

- Goal: characterize how populations adapt to nutrient limitation.
- Methods:
  1. Data: 24 populations of haploid and diploid Scer, under either glucose, phosphate or sulfate limitations for about 200 generations. Measure the change in gene expression pattern, and genomic sequences.
- Expression evolution:
  - The overall variations of expression across the same population (multiple clones) are higher in glucose or phosphate limitation than in sulfate limitation. Suggesting adapation under sulfate limitation is more constrained.
  - Metabolic adaptation strategies: Glu1 (some populations under glucose limitation) - downregulation of glycolysis; Glu2 - up-regulation of lipid oxidization and peroxisomal functions (thus may use lipid as energy source when glucose is limited) and Adh2 was up-regulated (ethanol consumption); Glu3 - Adh2 down-regulated.
- Genomic changes:
  - Genomic amplification of transporters: e.g. Hxt6/7 (high-affinity glucose transporters). Unequal mitotic recombination because of high similarity between Hxt6 and Hxt7.
  - Large duplications and rearrangements, transposons: play some role. E.g. insertion of Ty in Mth1, a negative regulator of glucose sensing. The effect is consitutive activation of Rgt1 leading to increased expression of sugar transporters.
  - SNPs: 34 SNPs in 10 clones/genomes. 84% in coding regions. Many coding mutations can be associated with known role in glucose metabolism, e.g. Ccr4, Mth1 and Snf6.
- Fitness effects of mutations: most are adaptive, with 5-10% fitness increase. One exception is Sul1 in sulfate limitation results in 50% fitness increase.

Evolution of bacterial chemotaxis network [Singh & Arkin, PNAS, 2008]

- Problem: how this network evolves in relation to the niche and life style?
- Methods:

  - Data: 207 bacterial species. Three networks considered: chemotaxis, sporulation and DNA uptake. The phenotypes are annotated as: non-motile, free-living, animal pathogen, plant pathogen, etc.
  - Evolutionary pattern: gene content in each species, and do clustering of genes on their phylogenetic profiles.

- Results:
  - Chemotaxis network evolves in modular fashion (genes gained or lost as modules): the modules correspond to the functions - sensing, regulation and actuction. Among all, actuation modules are more coherent and the other modules evolve faster.
  - Similar pattern holds for sporulation, but not DNA uptake.
  - Correlation with phenotypes: modules can be predictive of phenotypes, but there are exceptions, e.g. non-motile genes may contain actuators (flagellar) - because the orthologs of flagellar system are type III secretion system. In some motile species, the actuaors switched to type II system.

- Discussion: module rewiring as a way of adaptation:
  - Actuator modules may evolve multiple times: from different types of secretion systems.
  - Sensor apparatus seems to have evolved once, with niche-specific gene loss, and the transcriptional regulators seem to be recruited from other pathways.

Speciation of mangroves [Ziwen He and Suhua Shi, MicroEvol meeting, 2018]. Speciation with gene flow via cycles of isolation and migration: insights from multiple mangrove taxa [NSR, 2019]

- Mangrove: tidal wave, high salt. Adaptation strategies: secrete salt, root can breath.
- Experiment: studying speciation of R. Mucronata and R. Stylosa, two species that have morphological difference. For each species, multiple populations.
- Background: modes of speciation, allopatric (geographic isolation), sympatric (same place) and parapatric (adjacent).
- Background: introgression. Hybrid backcross repeatedly with one parental species, introducing gene flow to a receiving population.
- Phylogeny: (1) NJ tree: M1 (one type of M) and S clustered. (2) ML tree: as expected. (3) Chloroplast tree (based on maternal inheritance): similar to NJ tree. The discordance between NJ, ML and cholorplast trees can be explained by gene flow between M1 population and the RS species.
- Discovering highly divergent regions or genomic islands between two species: similar within species, but divergent across species. Several islands are found, enriched with genes with possible adaptive functions (stress, salt, etc.)
- Problem: the two species are geographically close, and gene flows are expected, how did speciation occur? Time needed for speciation: 1M years.
- PSMC analysis of population size changes, and correlation with climate changes.
- Gene flow analysis: similar to admixture, estimation proportion of ancestry. Genetic diff. of two species in Hainan (the regions are different in temperature, humidity).
- Model: climate fluctuation, esp. ice ages, reduces sea level and restrict gene flows. Repeated episodes of reduced gene flow can lead to speciation. Each cycle may be shorter than 1M years.

- Lesson: a general question is how speciation is maintained, especially when there is gene flow. One possible answer is the repeated disruption of gene flow. Also may explain the diversity of cichlid fish at Lake Victoria.
- Lesson: useful genetic analysis for speciation may include (1) Gene flow between the two species, including introgression (any regions in one species that contain genes of the other). (2) Divergence analysis: finding genomic islands that differ between species. (3) PSMC: reconstruct the population history, and comparison with environmental changes.

4.2 Microevolution

Can genomics shed light on the origin of species? [PLoS Bio, 2019]

- Questions about speciation: (1) identify speciation genes, and related questions, coding or regulatory, de novo or standing. (2) Broader questions: geographical scenarios.
- Possible strategy: pattern of F_{ST} across multiple lineages. In such comparison, often find strong parallelism, the same pattern of variation across the genome. This pattern can be created by multiple selection scenarios.
- Model 1: Background selection: strongest in high gene density and low recombination rate regions. However, may not be strong enough.
- Model 2: Repeated positive selection: lead to reduced diversity in the same regions across lineages. This is possible especially because most variants are standing.
- Model 3: Patterns of genetic variation may predate the common origin of multiple species, such that apparently independent lineages in a species group may not be truly independent of one another.

How islands shrink people [Science, 2018]

- Data: 32 pygmies of Flores, genotyping.
- Population ancestry: unrelated to Hobbit, rather East Asians.
- Positive selection in a gene involved in fat metabolism (consumption of sea food). Polygenic selection of height (short stature).

Natural Selection Has Differentiated the Progesterone Receptor among Human Populations [AJHG, 2018]

- PGR: positive selection in human, fixation in East Asian, but polymorphic in Europe. The fixed allele is associated with PTB.

Denisovan, modern human and mouse TNFAIP3 alleles tune A20 phosphorylation and immunity [NI, 2019]

- TNFAIP3 allele found in a AID family. Shown experimentally that it leads to stronger immune response.
- Tracing the history of the rare allele: only in oceania people, but not elsewhere. Also in a Denisovan girl. Show evidence of positive selection.
- Remark: an example of positive selection acting on immune genes during human evolution.