Ein Algorithmus zur Identifikation von Änderungen im Selektionsdruck

Hintergrund

Bei proteinkodierenden Sequenzen kann Selektion durch ein $\omega > 0$ parametrisiert werden – das Verhältnis zwischen nichtsynonymer und synonymer Mutationsrate ($\omega = \mathrm{dN/dS}$). Ein $\omega < 1$ bedeutet dabei, dass Mutationen von Kodons bevorzugt werden, die die Aminosäure nicht ändern. Dies wird auch als "negative" oder "reinigende" Selektion bezeichnet und wird durch die Redundanz des genetischen Codes ermöglicht. Ein $\omega \approx 1$ bedeutet neutrale Evolution. Dies kann etwa vorkommen, wenn eine Sequenz gar nicht (mehr) proteinkodierend und funktional ist. Ein $\omega > 1$ bedeutet, dass (bestimmte) Änderungen der Proteinsequenz bevorzugt in der Population fixiert werden, etwa weil der Organismus, in dem das Protein vorkommt, auf geänderte Umweltbedingungen reagieren muss und eine Veränderung des Proteins seine Fitness erhöht (z.B. "immune evasion" bei Viren). Dieses nennt man "positive" Selektion.
Methoden, solch ein ω zu schätzen, liefern in der Regel nur eine Schätzung $\hat{\omega}$ für die Eingabe:

- ein multiples Alignment von Kodonsequenzen und
- ein gewurzelter phylogenetischer Baum $T = (V, E)$ mit Zweiglängen für die Eingabesequenzen.

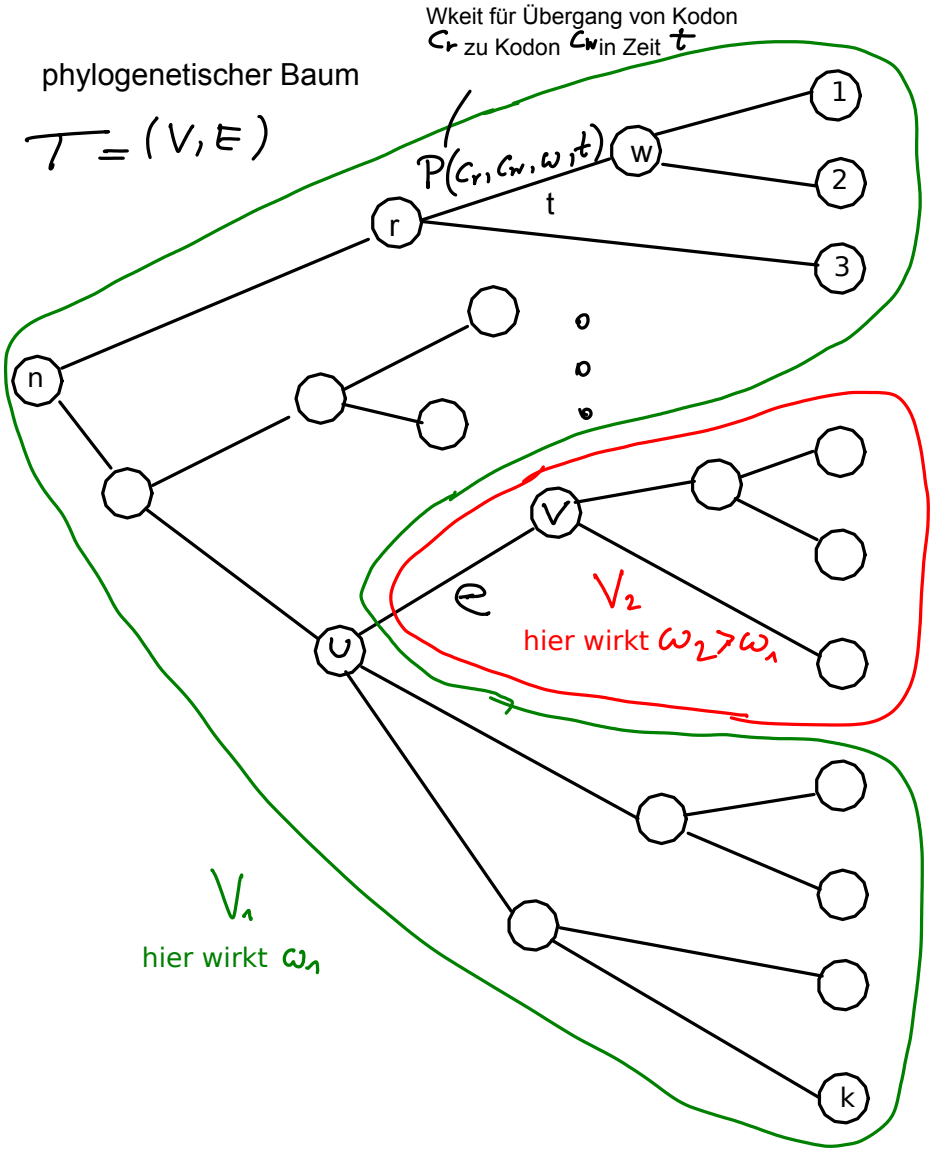
Sei $C = \{0, 1, \ldots, 63\}$ die Menge der Kodons. Dabei werden die Kondons typischerweise durchnummeriert: `aaa`, `aac`, `aag`, `aat`, ..., `ttt`. Für ein gegebenes ω und eine Kante (u, v) in einem phylogenetischen Baum mit Länge $t > 0$, bezeichne

$$P(c_u, c_v, \omega, t)$$

die Wahrscheinlichkeit, dass sich ein Kodon c_u nach Zeit t zu einem Kodon c_v verändert, wenn die Evolution mit Parameter ω abläuft, also dass an Knoten v das Kodon $c_v \in C$ steht, wenn an Knoten u das Kodon $c_u \in C$ steht. In der Regel ist $P(i, j, \omega, t)$ am größten für $j = i$ und für $j \neq i$ wächst $P(i, j, \omega, t)$ monoton in einer Umgebung von $t = 0$, die viele tatsächlich vorkommenden Astlängen enthält. Die $P(i, j, \omega, t)$ werden mit einer Markowkette stetiger Zeit modelliert und stehen in der AG Bioinformatik bereits als C++-Code bereit.

Aufgabe

- Entwickle die Details zu einem effizienten Algorithmus, der das Problem löst, ω in einem phylogenetischen Baum zu schätzen, wenn erlaubt ist, dass ω sich in maximal einem Teilbaum geändert hat, und sonst konstant ist für alle Spalten (Sites) und Kanten.
- Implementiere den Algorithmus und wende ihn auf Maus-Anwendungsdaten an.
- Analysiere und diskutiere die Ergebnisse.



Kodon Alignment

1 2 3 . . . e Spalten

cct	cca	ccg		cca
cct	cca	---		cca
cca	cca	cca		ccg

genetischer Code:
Prolin wird von den
4 Kodons cc{a,c,g,t} kodiert

cca	cca	cct		cta
tca	cca	cca		cca
cca	cga	cga		cca
cca	ccg	cca	. . .	cca
cca	---	cca		ccg
ccc	cca	cta		cca
cca	cca	ccc		cca

hier mehr nicht-synonyme Substitutionen ( )

wahrscheinl. eine synonyme Substitution

Sei $V = \{1, 2, \ldots, k, k+1, \ldots, n\}$, wobei $1, \ldots, k$ die den k Eingabesequenzen entsprechenden Blätter seien und $k+1, \ldots, n$ die inneren Knoten (n die Wurzel). Das multiple Kondonalignment sei gegeben durch

$$(c_{ij})_{\substack{1 \le i \le k \\ 1 \le j \le \ell}},$$

wobei $c_{ij} \in C$ das j-te Kodon der i-ten Sequenz sei (c_{ij} kann fehlen bei einer Alignmentlücke). Für einen Split der Baumknoten $V = V_1 \dot\cup V_2$ entlang einer Kante e sei a, die Funktion, die einem Knoten $w \in V$ seinen Teil, 1 oder 2, zuordnet: $a(w) = 1$, falls $w \in V_1$ und $a(w) = 2$, falls $w \in V_2$. Gesucht ist ein solcher Split und ω_1, ω_2, so dass

$$L(e, \omega_1, \omega_2) = \prod_{j=1}^{\ell} \sum_{\substack{c_{ij} \in C, \\ k < i \le n}} \prod_{(r,w) \in E} P(c_{rj}, c_{wj}, \omega_{a(w)}, t(w)) \tag{1}$$

maximiert wird. Dabei sei $t(w)$ die gegebene Länge der Kante (von r) zu w. Die Summation geht über alle Belegungen der $n-k$ *inneren* Knoten mit (anzestralen) Kodons ($i > k$), da die Blätter bereits mit Kodonsequenzen belegt sind (c_{ij} gegeben für $i \le k$). Um eine *a-priori* Verteilung der Kodons zu berücksichtigen, bildet die Wurzel eine Ausnahme. Wir nehmen an, dass auch die Wurzel $w = n$ einen Vorfahren 0 hat und dass $P(0, n, \omega, t) = \pi_n$ ist für einen vorgegebenen Vektor $\pi = (\pi_0, \ldots, \pi_{63})$ von Kodonwahrscheinlichkeiten.

Eingabedaten und Anwendungsbeispiel

Ein Datensatz zum Anwenden und Analysieren stammt aus einem Projekt zur Annotation von 18 neuen Mausstämmen. Hier hat Stefanie König mehrere tausend intronlose Genkandidaten gefunden, die auf Plausibilität getestet werden sollen, da der Verdacht besteht, dass viele von diesen Sequenzen sogenannte *Pseudogene* sind. Dies sind Sequenzen, die denen eines proteinkodierenden Gens sehr ähneln, aber, die keiner Funktion unterliegen, also tatsächlich kein Gen sind. Solche Pseudogene können entstehen, wenn richtige Gene kopiert und an anderer Stelle in das Genom eingefügt werden – in der Regel ohne Introns. Die zu überprüfende Hypothese ist, dass das ω sich entlang der Kante, die zu dem Teilbaum (Clade) mit den Pseudogenen führt von einem Wert deutlich kleiner als 1 zu einem Wert nahe bei 1 ändert. Dies würde dann einen neuen Ansatz liefern, um Pseudogene zu identifizieren und umgekehrt intronlose Gene, die sehr wahrscheinlich echt sind ($\omega \ll 1$ im ganzen Baum). Bei einzelnen Testdaten, ist bekannt, welche Sequenzen Pseudogene sind, und welche nicht. Diese können zum Testen der Hypothese benutzt werden.

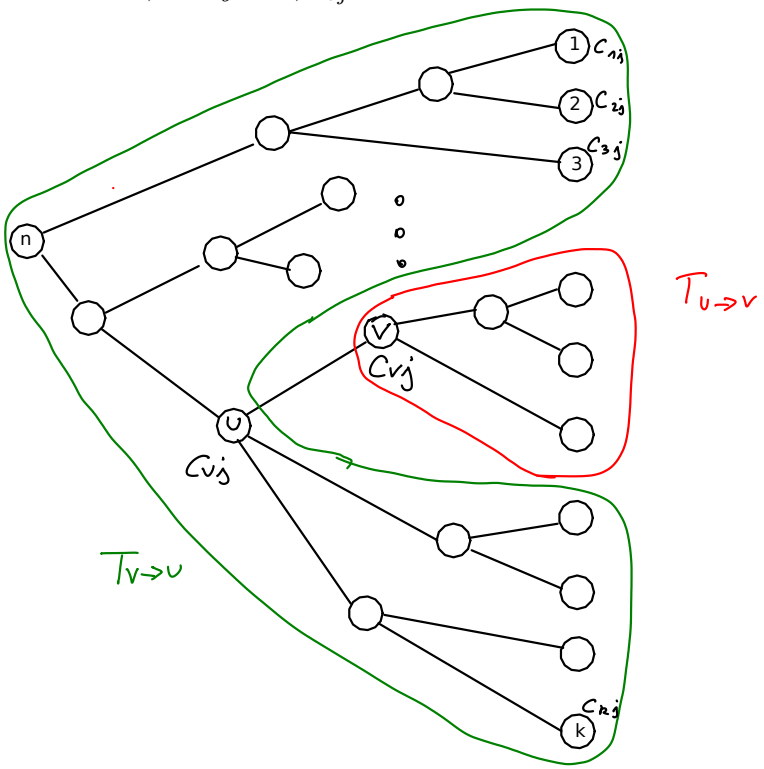
Mögliche Herangehensweise

Eine effiziente Lösung kann mit dynamischem Programmieren entlang des Baumes gefunden werden. Dabei kann davon ausgegangen werden, dass es eine kleine Menge $\Omega = \{q_1, \ldots, q_m\}$ von in Frage kommenden Werten für ω_1, ω_2 gibt.
Betrachte ein Paar (u, v) von benachbarten Knoten in T, d.h. $(u, v) \in E$ oder $(v, u) \in E$. Bezeichne $T_{u \to v} = (V_{u \to v}, E_{u \to v})$ den Teilbaum von T, der auf derselben Seite dieser Kante

liegt wie v (siehe Zeichnung unten). Betrachte eine einzelne Spalte j. Definiere die DP-Variablen

$$S(u, v, j, c_{vj}, \omega) := \sum_{\substack{c_{ij} \in C, \\ k < i \leq n, \\ i \in V_{u \to v} \setminus \{v\}}} \prod_{(r,w) \in E_{u \to v}} P(c_{rj}, c_{wj}, \omega, t(w))$$

für alle adjazenten u und v, $1 \leq j \leq \ell$, $c_{vj} \in C$ und $\omega \in \Omega$.



Beobachtungen:

- Die DP-Variablen $S(u, v, j, c_{vj}, \omega)$ lassen sich alle rekursiv in einem Hoch- und Runter-Durchlaufen von T in Linearzeit berechnen.
- Wenn alle $S(u, v, j, c_v, \omega)$ berechnet sind, lassen sich daraus das gesuchte Maximum, die entsprechende Kante e und ω_1, ω_2 ermitteln.

Voraussetzungen:
C++, Perl, dynamisches Programmieren, Satz von Bayes. Ggf. anlesen: Kodon-Substitutionsmodell mit ω. Matrix-Exponential e^Q

Personen: Lizzy Gerischer, Mario Stanke, Stefanie König

Literatur:

- Ziheng Yang, "Computational Molecular Evolution"
- Folienskript von M. Stanke zu "Molekulare Evolution", `http://www.math-inf.uni-greifswald.de/images/Stanke/molevo11/protevo.pdf`, `http://www.math-inf.uni-greifswald.de/images/Stanke/molevo11/treeprob.pdf`