

## Påvisning og karakterisering av positivt seleksjonstrykk i proteinkodende gener

av David A. Liberles

**En organismes genom har opp gjennom evolusjonen blitt formet av mange prosesser. Genduplikasjon, som kan skyldes rekombinasjon, transposisjon, polymerase-feil og andre mekanismer, former genomenes gener og funksjonelle elementer. Duplikasjonene blir i sin tur påvirket av setespesifikke mutasjoner, inersjons/delesjonsmutasjoner (indels), drift, omstokking av funksjonelle enheter, og til slutt seleksjon. Gjennom slike prosesser kan en få endringer i gensekvensen, strukturforandringer i proteinene, endringer i spleise-mønstre og i de regulatoriske forhold som gjør at ekspresjonsmønstrene kan forandre seg**

Interessante tilfeller av indel-begivenheter i samband med seleksjon har blitt oppdaget (Podlaha and Zhang, 2003). I denne artikkelen ble lengden av linkerene som regulerer N-terminal inaktivering av spenningsregulerte kalsiumkanaler vist å være under positivt seleksjonstrykk mellom ulike primat-ortologer.

Punktmutasjoner er den best studerte kilden til positiv seleksjon. Punktmutasjoner er en tilfeldig prosess, og mutasjoner drifter gjennom en populasjon avhengig av populasjonsstørrelse og –struktur, noe som i siste instans resulterer i en fiksering eller tap av mutasjonen i populasjonen. Seleksjonen kan inntreffe på mange nivåer, fra genomisk seleksjon for GC-innhold, til mRNA-seleksjon for baseparring, til translasjonell seleksjon for enkelte kodons basert på ulike tRNA-molekyler. Seleksjon på proteinnivået inntreffer på toppen av de ovenfor nevnte typer.

Seleksjonstrykk på proteinnivået som avviker fra nøytralitet kan være enten negativt eller positivt. Et negativt seleksjonstrykk inntreffer når en aminosyre er optimalt tilpasset dens potensielle funksjonelle rolle i proteinet. I dette tilfellet blir enhver annen aminosyre eliminert fra populasjonen fordi organismene blir mindre tilpasset. I praksis vil det eksistere et kontinuum av seleksjon mellom sterkt negative (bare en aminosyre er forenlig med funksjon), til negativ (for eksempel der bare fem hydrofile aminosyrer som KREND er forenlig med funksjon) og til nøytral (alle de tjue aminosyrene er like vel forenlige med funksjon).

Et positivt seleksjonstrykk inntreffer når en organisme med en mutasjon har en høyere "fitness" enn de uten, og når mutasjonen blir fiksert i populasjonen gjennom drift. Aminosyren under positivt seleksjonstrykk kan representere en nyvinning som i siste instans tillater organismen å tilpasse seg sitt levested på en ny mate. Substitusjonen kan også være et svar på en svakt skadelig substitusjon ellers i proteinet, eller i et interagerende protein

En generell observasjon, som har blitt gjort, og som er uavhengig av proteinfunksjon, er at enkeltkopigener (de gener som har ortologe likheter til nære arter utenom egen gruppe) og nylig dupliserte gener i et genom (de som viser paraloge forhold til nære arter utenom egen

gruppe) viser ulik oppførsel (se for eksempel Wagner, 2000). Ohno (Ohno, 1970) foreslo i et meget berømt arbeid at genduplikasjoner var en viktig mekanisme for å skape evolusjonære nyheter. Mekanismen bak dette var at genduplikasjoner ble frie fra negativt seleksjonstrykk, noe som satte dem i stand til å utforske det mulige sekvensrommet inntil en av kopiene avvek tilstrekkelig fra utgangspunktet til ikke lenger å kunne utføre sekvensens opprinnelige funksjon. Denne kopien ble dermed frigjort til å fortsette å utforske sekvensrommet med sikte på finne nye funksjoner, eventuelt med positiv seleksjon og neofunksjonalisering. Lynch (Force et al., 1999) har karakterisert neofunksjonalisering, subfunksjonalisering og pseudogenisering som alternative sluttpunkter for dupliserte gener.

Med en interesse for å forstå neofunksjonalisering langs enkeltlinjer som utgangspunkt, skal vi nå se litt på metodologi for å påvise den. Den genetiske koden inneholder redundans i enkelte posisjoner, ved at flere koder kan gi samme aminosyre. Disse posisjonene, oftest tredje-posisjon i kodons, kalles synonyme, og deres substitusjonsrate kalles den synonyme substitusjonsrate. Posisjoner der substitusjoner kan endre de kodede aminosyrer kalles ikkesynonyme posisjoner, og deres substitusjonsrate kalles den ikkesynonyme nukleotidsubstitusjonsrate. Ratioen mellom ikkesynonym og synonym substitusjonsrate (også kjent som  $Ka/Ks$ ,  $dN/dS$  eller  $\omega$ ) er en indikator på seleksjonstrykk i et gen. Når  $Ka$  er signifikant større enn  $Ks$  er dette evidens for positiv seleksjon, der individene med aminosyremutasjoner er bedre tilpasset enn individer uten slike mutasjoner. I forbindelse med en "multiple sequence alignment" og fylogenetiske trær kan  $Ka/Ks$  tilpasses som en grein-spesifikk parameter innen en sannsynlighetsmodell (Yang, 1998). Alternativt kan definerte ursekvensers avvik ved forgreningspunktet bestemmes, og  $Ka/Ks$  bestemmes parvis mellom sekvensene lang greinene (Messier and Stewart, 1997).

Endelig kan en gjennomsnittsberegning av  $Ka/Ks$  over hele genet oppfattes som et konservativt mål for påvisning av positivt seleksjonstrykk. I et protein som er gjenstand for neofunksjonalisering vil posisjonene som påvirker proteinfolding sannsynligvis være under

negativt seleksjonstrykk, mens de som for eksempel i en proteinbindende lomme er under utvikling for å binde en ny ligand, vil være under et positivt seleksjonstrykk. En metode for å dele opp sekvensen kalles primære sekvensvinduer, der Ka/Ks beregnes over en rekke av sammenhengende aminosyrer i den primære sekvensen (Endo et al., 1996; Fares et al., 2002). En annen metode er å dele opp sekvensen i de som er helt uforanderlige gjennom evolusjonshistorien, og de som har vist variasjon. Deretter beregnes Ka/Ks for de aminosyrene som varierer (Siltberg and Liberles, 2002). I siste instans er det struktur og funksjon som er gjenstand for seleksjon. En tredje metode, kalt tertiære sekvensvinduer, beregner Ka/Ks i et sirkulært rom med en gitt radius rundt hver aminosyreposisjon (Berglund and Liberles, under arbeid).

En annen innfallsvinkel for å påvise positivt seleksjonstrykk er å se etter skift i substitusjonsmodellen. En spesifikk sannsynlighetsratiotest er utviklet for å øke antall parametre og å etterspørre en tilhørende signifikant forbedring av sannsynlighetsscore. Den mest vanlige parameteren som benyttes, kalt  $\alpha$ , er formparameteren av  $\Gamma$ -distribusjonen av aminosyrerater på tvers av posisjoner (Gaucher et al., 2002). Den optimale substitusjonsmatris som beskriver en aminosyre kan også testes (Soyer et al., 2003).

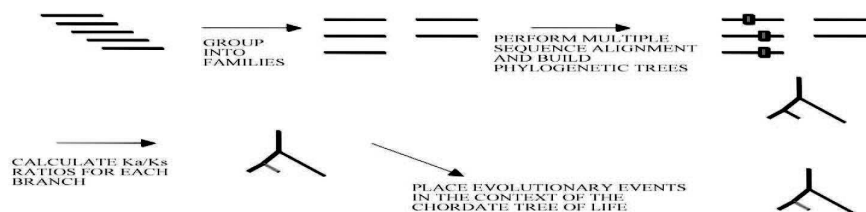
Disse metodene kan også anvendes for å oppdage neofunksjonalisering i stor skala i kordaters genfamilier. Vår første innfallsvinkel for å påvise neofunksjonalisering brukte en "Master Catalog" som utgangspunkt (Liberles et al., 2001). I senere tid har vi begynt å bruke våre egne genfamilier, tilgjengelige på <http://www.bioinfo.no>.

En enkel prosedyre for å bygge gode genfamilier er å starte med et alle-mot-alle BLAST-søk for å eliminere proteiner som ikke er nært beslektet med hverandre. Grenseverdien en bruker for slektskap kan defineres avhengig av hvorvidt familiedefinisjonen er basert på artstaksonomi eller sekvensavstand. Hvis en er interessert i sekvensavstand som kriterium, kan man filtrere BLAST-treff med en definert PAM-avstand og parvise lengdeterskler. Ulike algoritmer er tilgjengelige

for enkel og fullstendig "linkage clustering" for å danne familier. Når et familiesett har blitt identifisert, kan proteinbaserte multiple sequence alignments og fylogenetiske trær beregnes. Den presise prosedyre avhenger av de vitenskapelige spørsmål som skal besvares. Den generelle prosedyren, utvidet for å påvise positiv seleksjon i fylogenetisk sammenheng, er vist i figuren nedenfor.

Ved en beregning av Ka/Ks over alle 5305 kordatgenfamilier i vår Master Catalog, kunne vi påvise 643 greiner fra 280 genfamilier som representerte 63 greiner i NCBI-taksonomien (Benson et al., 2004) ved hjelp av en genre-til-artstre-kartlegging (Liberles et al., 2001). Disse ble samlet inn i en database kalt "The Adaptive Evolution Database", eller TAED. Etter som stadig flere kordat-genomer sekvenseres vil genfamiliene bli stadig mer presise, med stadig kortere grein-avstand mellom genene. Mange av de positivt selekterte genene i den opprinnelige TAED-beregningen var av immunsystem-gener og gener involvert i reproduksjon, dvs. to klassiske eksempler på det som kan kalles evolusjons-våpenkappløpet. To pattedyrgener vi undersøkte i stor detalj, basert på den opprinnelige TAED-beregningen, var leptin og plasminogen aktivator.

Leptin har blitt identifisert som et fedme-gen i mus. Leptin-manglende mus er fete, og behandling med leptin eliminerer fedmen. Imidlertid synes mus som et dårlig modellsystem for leptin i mennesker (Benner et al., 1998). Leptin-treet i TAED antyder en primat-spesifikk evolusjon. The kunne også sees som et skift i  $\alpha$  mellom primater og andre pattedyr av Eutheria i leptingenfamilien (Siltberg and Liberles, 2002). Videre ble det påvist at leptinreseptorens ekstracellulære del viste tilsvarende mønster for positiv seleksjon for leptin (Benner et al., 1998). Evidens for transposon-mediert nyekspresjon av leptin i morkake og en annen inserjonsbegivenhet i leptins reseptor gjennom primatevolusjonen har også kommet til syne (Bi et al., 1997; Kapitonov and Jurka, 1999). Ved hjelp av strukturinformasjon har et nytt bindingssete i leptin blitt foreslått (Gaucher et al., 2003). Videre har bindingsflaten mellom leptin og reseptoren



*En skjematisk framstilling av databasert funksjonsgenomisk analyse av genfamilier, via multipl sekvenssammenstilling og fylogenetiske trær og bruk med evolusjonshypoteser som Ka/Ks. I siste instans er slike malinger korrelert til andre begivenheter som inntreffer i samme del av livets tre..*

blitt gjenstand for radikale substitusjoer, som vist i datamaskin-modeller (Hiroike et al., 2000). Alt dette peker mot neofunksjonalisering av leptin i primater, og en delvis forskjellig funksjonell rolle sammenliknet med den i mus.

Plasminogen aktivator har vært utsatt for multiple genduplikasjoner mellom *Desmodus rotundus* (vanlig vampyrflaggermus) og den nærmeste ikke-blodsugende slektning, *Carollia perspicillata* (kalifornisk korthalet fruktflaggermus) (Kemi, Savolainen, and Liberles, upubliserte observasjoner). Positiv seleksjon er tydelig i TAED fra den bovine utgruppen av genduplikasjoner, så vel som langs greinene som skiller duplikasjonene fra hverandre (Liberles et al., 2001). Vampyrflaggermus-paralogene, som uttrykkes i spytt, gjør vampyrflaggermusene i stand til å forhindre blodklumping ved suging av blod fra et bitt. Pågående sekvensering og analyse har som mål å forstå de ulike funksjonelle roller for de ulike duplikater. Fire paraloger har blitt sekvensert fra vanlige vampyrflaggermus, men ingen med to kringel-domener, som er nødvendig for bovin regulering av plasminogen aktivator. Videre er ingen sekvenser kjent for "harry-legged" vampyrflaggermus (*Diphylla ecaudata*) eller hvit-vinget vampyrflaggermus (*Diaemus youngi*). Det innebærer at tidsrekkefølgen av duplikasjonsbegivenhetene, såvel som graden av konservering, foreløpig er uklart. Disse ubesvarte spørsmål er gjenstand for undersøkelse i pågående forskning i Bergen.

Ettersom gen- og genomsekvensering i ulike kordatarter øker, vil komparativ genomikk gi stadig større kraft til å påvise positiv seleksjon og neofunksjonalisering, for på den måten å gi hint om molekylære mekanismer for artsavhengig tilpasning. Samtidig utvikler metodene for å påvise slike forandringer seg. Denne kombinasjonen muliggjør et rammeverk for en bedre forståelse av kordatgenomenens evolusjon.

*Noter: Norsk oversettelse ved redaksjonen. Deler av denne artikkelen likner Liberles, D.A. "Detecting and Characterizing Adaptive Evolution in Chordate Proteins." in Comparative Genomics of Vertebrates: Concepts and Bioinformatic Tools. INSERM, 2004.*

#### Referanser:

- Benner SA, Trabesinger N, and Scheiber D. 1998. Post-genomic science: Converting primary sequence into physiological function. *Adv. Enzyme Regul.* 38:155-190.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. 2004. GenBank: Update. *Nucl. Acids Res.* 32:D23-D26.
- Bi S, Garilova O, Gong DW, Mason MM, and Reitman M. 1997. Identification of a placental enhancer for the human leptin gene. *J. Biol. Chem.* 272:30583-30588.
- Endo T, Ieko K, and Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13:685-690.
- Fares MA, Elena SF, Ortiz J, Moya A, and Barrio E. 2002. A sliding window-based method to detect selective constraints in protein coding genes and its application to RNA viruses. *J. Mol. Evol.* 55:509-521.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531-1545.
- Gaucher EA, Gu X, Miyamoto MM, and Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* 27:315-321.
- Gaucher EA, Miyamoto MM, and Benner SA. 2003. Evolutionary, structural, and biochemical evidence for a new interaction site of the leptin obesity protein. *Genetics* 163:1549-1553.
- Hiroike T, Higo J, Jingami H, and Toh, H. 2000. Homology modeling of human leptin/leptin receptor complex. *Biochem. Biophys. Res. Comm.* 275:154-158.
- Kapitonov VV and Jurka J. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *J. Mol. Evol.* 48:248-251.
- Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, and Benner SA. 2001. The Adaptive Evolution Database (TAED). *Genome Biol.* 2(8):research0028.1-research0028.6.
- Messier W and Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151-154.
- Ohno S. 1970. *Evolution by gene duplication.* New York: Springer-Verlag.
- Podlaha O and Zhang J. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc. Natl. Acad. Sci., USA* 100:12241-12246.
- Siltberg J and Liberles, DA. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. *J. Evol. Biol.* 15:588-594.
- Soyer OS, Dimmic MW, Neubig RR, and Goldstein RA. 2003. Dimerization in aminergic G-protein-coupled receptors: Application of a hidden-site class model of evolution. *Biochem.* 42:14522-14531.
- Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci., USA* 97:6579-6584.
- Yang ZH. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568-573.

#### Diagnose

Dersom du tror du bioinformatikk kan hjelpe deg i din forskning, men føler behov for rådgiving, oppfordrer vi deg til å ta kontakt med oss. Basert på en kort beskrivelse av problemstillingen tilbyr vi å stille en diagnose. Det vil si at vi basert på vår kompetanse vurderer hvilke bioinformatiske tilnæringer som kan være aktuelle, og hvor mye innsats som trengs for å gjennomføre eventuelle bioinformatiske analyser. Selve diagnosen er gratis!