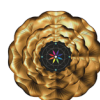


VLADIMIR CHERBAK

MAXIM MAKUKOV

I NUMERI DEL
CODICE GENETICO



Il segnale "Wow!" del codice genetico terrestre

Vladimir I. shCherbak^a e Maxim A. Makukov^{b*}

^aDepartment of Mathematics, al-Farabi Kazakh National University, Almaty,
Republic of Kazakhstan

e-mail: Vladimir.shCherbak@kaznu.kz

^bFesenkov Astrophysical Institute, Almaty, Republic of Kazakhstan

e-mail: makukov@gmail.com, makukov@aphi.kz

Questa è la versione degli autori del manoscritto pubblicato su Icarus.

Le modifiche derivanti dal processo di pubblicazione potrebbero non riflettersi in questo documento. Le modifiche possono essere state apportate a questo lavoro dopo l'invio per la pubblicazione. Per la versione della rivista si veda

<http://dx.doi.org/10.1016/j.icarus.2013.02.017>

Storia dell'articolo:

Inviato il 26 giugno 2012

Rivisto il 3 ottobre 2012

Rivisto il 31 gennaio 2013

Accettato il 12 febbraio 2013

Parole chiave: astrobiologia codice genetico panspermia diretta SETI

* Autore corrispondente

Per ulteriori informazioni vedere: <https://bioseti.info>

ABSTRACT

È stato ripetutamente proposto di ampliare il campo di applicazione del SETI e una delle alternative alla radio è rappresentata dai mezzi di comunicazione biologici. Il DNA genomico è già utilizzato sulla Terra per immagazzinare informazioni non biologiche. Il codice genetico è più piccolo in termini di capacità, ma più forte in termini di immunità al rumore. Il codice è una mappatura flessibile tra codoni e amminoacidi, e questa flessibilità permette di modificare il codice artificialmente. Ma una volta fissato, il codice può rimanere invariato su scale di tempo cosmologiche;

infatti, è il costrutto più duraturo che si conosca. Pertanto, rappresenta un deposito eccezionalmente affidabile per una firma intelligente, se questa è conforme ai requisiti biologici e termodinamici. Poiché lo scenario attuale dell'origine della vita terrestre è ben lungi dall'essere definito, non si può escludere la possibilità che sia stata seminata intenzionalmente. Un "segnale" intelligente statisticamente forte nel codice genetico è quindi una conseguenza testabile di tale scenario. In questa sede dimostriamo che il codice terrestre mostra un'accurata precisione e un ordine che soddisfa i criteri per essere considerato un segnale informativo. Semplici disposizioni del codice rivelano un insieme di schemi aritmetici e ideografici dello stesso linguaggio simbolico. Accurati e sistematici, questi schemi sottostanti appaiono come il prodotto di una logica di precisione e di un calcolo non banale piuttosto che di processi stocastici (l'ipotesi nulla che siano dovuti al caso accoppiato a presumibili percorsi evolutivi è respinta con un valore $P < 10^{-13}$). I modelli sono profondi al punto che la stessa mappatura del codice è dedotta in modo univoco dalla loro rappresentazione algebrica. Il segnale mostra segni di artificialità facilmente riconoscibili, tra cui il simbolo dello zero, la sintassi decimale privilegiata e le simmetrie semantiche. Inoltre, l'estrazione del segnale comporta operazioni logicamente semplici ma astratte, rendendo i modelli essenzialmente irriducibili a qualsiasi origine naturale. Vengono discussi i modi plausibili di incorporare il segnale nel codice e la possibile interpretazione del suo contenuto. Nel complesso, mentre il codice è quasi ottimizzato dal punto di vista biologico, la sua capacità limitata viene utilizzata in modo estremamente efficiente per memorizzare informazioni non biologiche.

Introduzione

Le recenti conquiste biotecnologiche consentono di utilizzare il DNA genomico come memoria di dati più duratura di qualsiasi supporto attualmente utilizzato (Bancroft et al., 2001; Yachie et al., 2008; Ailenberg & Rotstein, 2009). Forse l'applicazione più diretta è stata proposta ancor prima dell'avvento della biologia sintetica. Considerando i canali informativi alternativi per la SETI, Marx (1979) ha osservato che i genomi delle cellule viventi possono rappresentare un buon esempio. Ha anche osservato che il codice genetico è ancora più resistente. Esposto a una forte selezione negativa, il codice rimane invariato per miliardi di anni, salvo rari casi di variazioni minori (Knight et al., 2001) ed espansioni dipendenti dal contesto (Yuan et al., 2010). Eppure, la mappatura tra codoni e aminoacidi è malleabile, poiché essi interagiscono attraverso molecole modificabili di tRNA e aminoacil-tRNA sintetasi (Giegé et al., 1998; Ibba & Söll, 2000; vedi anche Appendice A). Questa capacità di riassegnare i codoni, che si pensa sia alla base dell'evoluzione del codice verso l'ottimizzazione a

più livelli (Bollenbach et al., 2007), permette anche di modificare il codice artificialmente (McClain & Foss, 1988; Budisa, 2006; Chin, 2012). È possibile, almeno in linea di principio, organizzare una mappatura che sia conforme ai requisiti funzionali e che contenga un piccolo messaggio o una firma, consentita dai 384 bit di capacità informativa del codice. Una volta che il genoma è stato opportunamente riscritto (Gibson et al., 2010), il nuovo codice con la firma rimarrà congelato nella cellula e nella sua progenie, che potrebbe poi essere consegnata attraverso lo spazio e il tempo a potenziali destinatari. Essendo efficiente dal punto di vista energetico (Rose & Wright, 2004) e autoreplicante, il canale biologico è anche libero dai problemi propri dei segnali radio: non è necessario fare affidamento sul tempo di arrivo, sulla frequenza e sulla direzione. A causa di queste restrizioni, l'origine del famoso segnale "Wow!" ricevuto nel 1977 rimane incerta (Ehman, 2011). Il canale biologico è stato preso in seria considerazione per i suoi meriti nel SETI, anche se con particolare attenzione ai genomi (Yokoo & Oshima, 1979; Freitas, 1983; Nakamura, 1986; Davies, 2010; Davies, 2012).

Nel frattempo, è stato proposto di assicurare la vita terrestre seminando esopianeti con cellule viventi (Mautner, 2000; Tepfer, 2008), e sembra che sia una questione di tempo. Il canale biologico si presta a questa impresa. Per evitare pregiudizi antropocentrici, si potrebbe ammettere che la vita terrestre non è il punto di partenza della serie di colonizzazioni cosmiche (Crick & Orgel, 1973; Crick, 1981). Se così fosse, è naturale aspettarsi un "segnale" intelligente statisticamente forte nel codice genetico terrestre (Marx, 1979). Questa possibilità è ulteriormente incentivata dal fatto che il modo in cui il codice è arrivato a essere apparentemente non casuale e quasi ottimizzato rimane ancora discutibile e altamente speculativo (per le recensioni sui modelli tradizionali di evoluzione del codice si veda Knight et al., 1999; Gusev & Schulze-Makuch, 2004; Di Giulio, 2005; Koonin & Novozhilov, 2009).

L'unico modo per estrarre un eventuale segnale dal codice è quello di organizzare i suoi elementi - codoni, amminoacidi e segni sintattici - in base ai loro parametri, utilizzando una logica semplice. Questi arrangiamenti vengono poi analizzati alla ricerca di schemi o strutture grammaticali di qualche tipo. La scelta delle disposizioni e dei parametri deve escludere l'arbitrarietà. Ad esempio, si dovrebbero considerare solo quei parametri che non dipendono da sistemi di unità fisiche. Tuttavia, anche in questo caso, a priori non si sa esattamente che tipo di modelli ci si possa aspettare. C'è quindi il rischio di falsi positivi, dato che con un insieme di dati come il codice genetico è facile trovare vari modelli di un tipo o dell'altro.

Tuttavia, il compito potrebbe essere in qualche modo alleggerito. In primo luogo, è possibile prevedere alcuni aspetti generali di un segnale putativo e del suo "linguaggio", soprattutto se si sfrutta l'esperienza attiva del SETI. Ad esempio, è generalmente accettato che il linguaggio numerico dell'aritmetica sia lo stesso per tutto l'universo (Freudenthal, 1960; Minsky, 1985). Inoltre, i simboli e la grammatica di questo linguaggio, come i sistemi numerici posizionali con concezione zero, sono segni distintivi dell'intelligenza. Così, i messaggi interstellari inviati dalla Terra di solito iniziano con una sequenza naturale di numeri in notazione binaria o decimale. Per rafforzare l'artificialità, nella posizione astratta che precedeva la sequenza veniva posto il simbolo dello zero. Questi messaggi includevano anche simboli di operazioni aritmetiche, triangolo egizio, DNA e altre nozioni di coscienza umana (Sagan et al., 1972; The Staff at the NAIC, 1975; Sagan et al., 1978; Dumas & Dutil, 2004). In secondo luogo, per ridurre al minimo il rischio di falsi positivi, si possono imporre requisiti il più possibile restrittivi a un segnale presunto.

Ad esempio, è ragionevole aspettarsi che un messaggio veramente intelligente rappresenti non solo un insieme di modelli di vario tipo, ma anche modelli dello stesso "stile linguistico". In questo caso, se si nota un potenziale modello, si può restringere la ricerca allo stesso tipo di modelli. Un altro requisito rigoroso potrebbe essere che i modelli devono coinvolgere ogni elemento del codice in ogni disposizione, mentre l'intero segnale deve occupare la maggior parte, se non la totalità, della capacità informativa del codice. In linea di massima, data la natura del compito, le specifiche della strategia vengono definite durante il percorso.

Seguendo queste linee, dimostriamo che il codice terrestre contiene un insieme di schemi di precisione che soddisfano i requisiti di cui sopra. La semplice sistematizzazione del codice rivela un forte segnale informativo che comprende componenti aritmetiche e ideografiche. È notevole che i modelli indipendenti del segnale siano tutti espressi in un linguaggio simbolico comune. Dimostriamo che il segnale è statisticamente significativo, impiega interamente la capacità informativa del codice e non è riconducibile a un'origine naturale. I modelli di comparsa della vita primordiale con un codice genetico originale privo di segnale esulano dallo scopo di questo articolo; qualunque cosa fosse, lo stato precedente del codice è stato cancellato dal palinsesto del segnale.

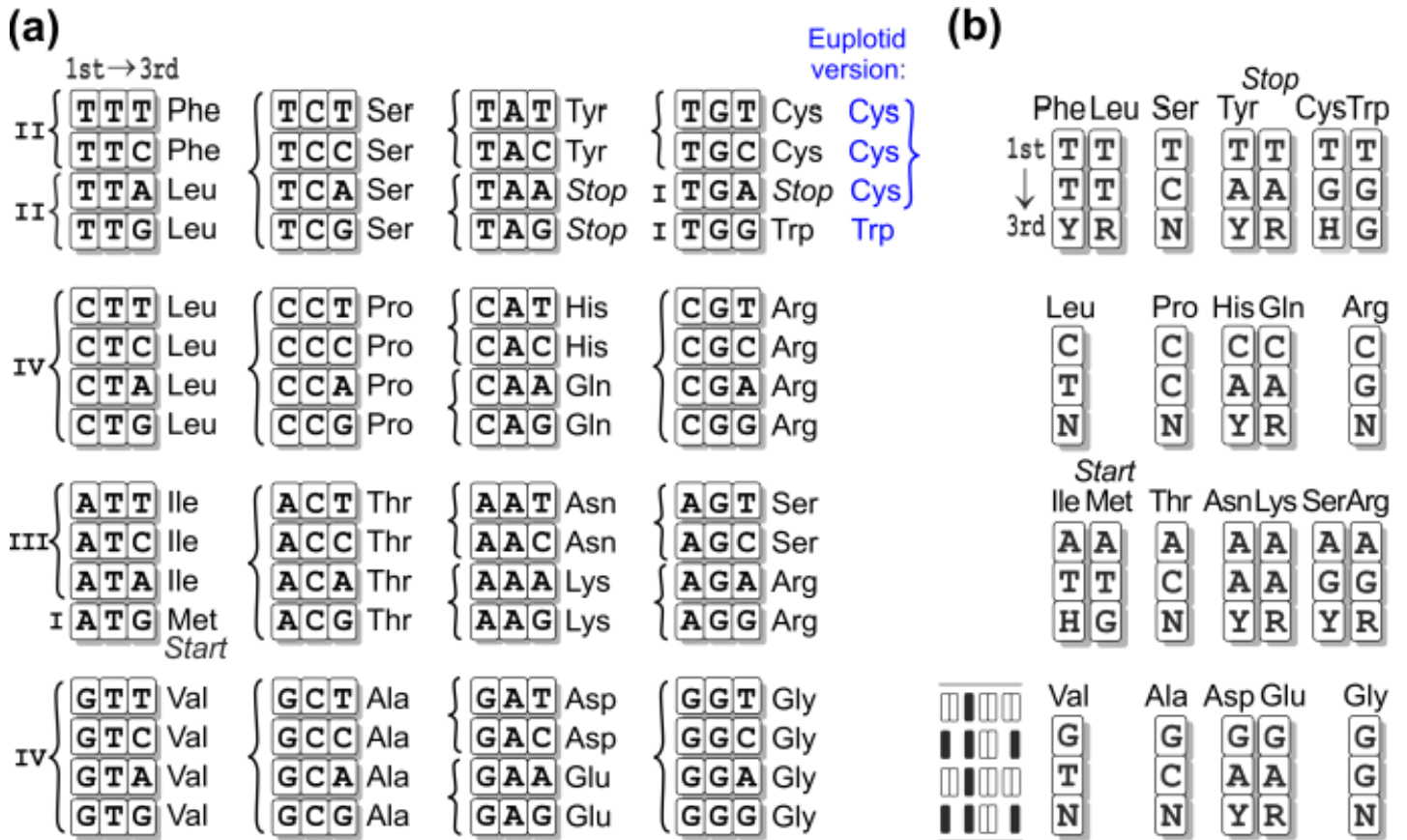


Fig. 1. Il codice genetico. (a) Rappresentazione tradizionale del codice standard, o universale. I codoni che codificano lo stesso amminoacido formano serie sinonimiche indicate con parentesi graffe di apertura. Il numero di codoni in una serie definisce la sua ridondanza (degenerazione). Intere famiglie di codoni sono costituite da una serie di ridondanza IV. Altre famiglie sono divise. La maggior parte delle famiglie divise sono dimezzate in due serie di ridondanza II ciascuna, una che termina con le pirimidine {T, C} e l'altra con le purine {A, G}. Tre codoni del codice standard non sono mappati su nessun amminoacido e sono usati come stop nella traduzione. L'inizio è solitamente indicato da ATG, che codifica Met. La parentesi graffa di chiusura mostra l'unica differenza tra il codice euplotidico e quello standard. (b) Rappresentazione contratta della versione euplotide. I codoni sinonimi a grandezza naturale sono sostituiti da una singola serie contratta con terza base combinata. Vengono utilizzate le denominazioni FASTA: R e Y stanno rispettivamente per purine e pirimidine, N per tutte e quattro le basi e H per {T, C, A}. Le serie sono disposte verticalmente per maggiore comodità. Il pittogramma a sinistra è utile per le figure sottostanti. Gli elementi riempiti indicano intere famiglie.

Il contesto

Se ci fosse un segnale nel codice, probabilmente si sarebbe manifestato in qualche modo durante il mezzo secolo di storia dell'analisi tradizionale dell'organizzazione del codice. È quindi utile riassumere brevemente ciò che è stato appreso al riguardo fino ad oggi. Inoltre, per semplificare la presentazione dei dati, citeremo in anticipo alcune informazioni a posteriori relative al segnale da descrivere, che verranno discusse in modo più approfondito a tempo debito. Sugeriamo al lettore che non ha familiarità con i meccanismi molecolari alla base del codice genetico di fare riferimento all'Appendice A, dove è spiegato anche perché il codice è suscettibile di "modulazione" intenzionale (per usare il linguaggio del SETI radio-orientato) e, allo stesso tempo, è altamente protetto dalla "modulazione" casuale (ha una forte immunità al rumore).

Il codice in sintesi. Non appena il codice genetico è stato decifrato biochimicamente (Nirenberg et al., 1965), la sua struttura non casuale è diventata evidente (Woese, 1965; Crick, 1968). Il modello più evidente emerso nel codice è la sua regolare ridondanza. Il codice comprende 16 famiglie di codoni che iniziano con la stessa coppia di basi e queste famiglie consistono generalmente in una o due serie uguali di codoni mappati su un amminoacido o su Stop (Fig. 1a). In effetti, il codice standard è quasi simmetrico in termini di ridondanza. Ci sono solo due famiglie divise in modo ineguale: quelle che iniziano con TG e AT. L'azione minima per ripristinare la simmetria è far coincidere la famiglia TG con la famiglia AT riassegnando TGA da Stop a cisteina. Per inciso, questa versione simmetrica non è solo un'ipotesi teorica, ma si trova anche in natura come codice nucleare dei ciliati euplotidi (Meyer et al., 1991). Mentre il codice standard memorizza la componente aritmetica del segnale, la versione simmetrica euplotidea conserva quella ideografica (l'interrelazione tra queste due versioni del codice è discussa più avanti). La ridondanza regolare porta anche alla struttura a blocchi del codice genetico. Ciò consente di rappresentare il codice in una forma contratta, in cui ogni amminoacido corrisponde a un singolo blocco, o in una serie contratta (Fig. 1b).

Le tre eccezioni sono Arg, Leu e Ser, che hanno una serie IV e una serie II ciascuna. Oltre alla ridondanza regolare, in seguito sono state riportate numerose altre caratteristiche, tra cui la robustezza agli errori (Alff-Steinberger, 1969), la correlazione tra termostabilità e ridondanza delle famiglie di codoni (Lagerkvist, 1978), la distribuzione non casuale degli amminoacidi tra i codoni se giudicata in base alla loro polarità e voluminosità (Jungck, 1978), le vie biosintetiche (Taylor & Coates, 1989), la reattività (Siemion & Stefanowicz, 1992) e persino il gusto (Zhuravlev, 2002). È stato anche dimostrato che il codice è efficace nel gestire informazioni aggiuntive nel DNA

(Baisnée et al., 2001; Itzkovitz & Alon, 2007). Apparentemente, queste caratteristiche sono legate, semmai, alla funzione biologica diretta del codice. Esistono anche diversi approcci astratti al codice, come quelli basati sulla topologia (Karasev & Stefanov, 2001), sulla scienza dell'informazione (Alvager et al., 1989) e sulla teoria dei numeri (Dragovich, 2012). Tuttavia, l'obiettivo principale di questi approcci è la costruzione di modelli teorici di descrizione di caratteristiche note del codice, piuttosto che affrontarne di nuove.

Nel complesso, solo due regolarità intrinseche, osservate all'inizio dello studio del codice, potrebbero suggerire una possibile relazione con un segnale putativo, grazie al loro carattere evidente e non ambiguo. Suggestiscono inoltre due parametri interi adimensionali per l'estrazione del segnale. Si tratta della quantità di codoni in una serie mappata su un amminoacido (ridondanza) e della quantità di nucleoni nelle molecole di amminoacidi. Questi parametri potrebbero essere chiamati "numeri ostensivi" per analogia con la quantità di segnali radiofonici in *Lingua Cosmica* (Freudenthal, 1960).

La bisezione di Rumer. Rumer (1966) ha suddiviso il codice per ridondanza - il primo "numerale ostensivo". Nel codice ci sono 8 famiglie intere e 8 famiglie divise (Fig. 2a). Rumer scoprì che i codoni di queste famiglie sono mappati l'uno sull'altro in modo uno a uno con una semplice relazione TG, CA, oggi nota come trasformazione di Rumer. Esistono altre due trasformazioni di questo tipo: TC, AG e TA, CG. Anch'esse compaiono nella bisezione di Rumer e ciascuna rende la metà di quanto rende la sola trasformazione di Rumer. Una bisezione arbitraria del codice ha poche possibilità di produrre una trasformazione, e ancora meno - il loro insieme ordinato (vedi Appendice B). La scoperta di Rumer è stata riscoperta da Danckwerts e Neubert (1975), che hanno anche notato che questo insieme potrebbe essere descritto con una struttura nota in matematica come gruppo di Klein-4. Questo ha dato il via a una serie di altre scoperte. Ciò ha dato il via a una serie di altri modelli che hanno coinvolto la teoria dei gruppi per descrivere il codice (Bertman & Jungck, 1979; Hornos & Hornos, 1993; Bashford et al., 1998), i quali, per la verità, non hanno ottenuto risultati decisivi. Nel frattempo, nelle teorie tradizionali sull'evoluzione del codice questa caratteristica è stata del tutto ignorata, sebbene sia stata ripetutamente riscoperta (ad esempio, si veda Wilhelm & Nikolajewa, 2004). È da notare che questa regolarità - che risulta essere una piccola porzione del segnale - è stata notata per la prima volta subito dopo la delucidazione dell'assegnazione dei codoni. Insieme al fatto delle riscoperte, ciò dimostra la natura anticrittografica del segnale all'interno del codice.

I nucleoni degli amminoacidi. Hasegawa & Miyata (1980) hanno disposto gli amminoacidi in ordine crescente di numero di nucleoni - il secondo "numero ostensivo" che, a differenza di altre proprietà degli amminoacidi, non si basa su un sistema di unità scelto arbitrariamente. Tale disposizione rivela un'approssimativa anticorrelazione: maggiore è la ridondanza, minore è il numero di nucleoni (Fig. 2b). Ciò ha fatto ipotizzare che gli amminoacidi piccoli prevalenti abbiano occupato le serie di maggiore ridondanza durante l'evoluzione del codice. Come mostrato di seguito, questa anticorrelazione è una derivata del segnale. Inoltre, proprio questa osservazione suggerisce una semplice sistematizzazione per entrambi i "numeri ostensivi": la disposizione monotona dei nucleoni e dei numeri di ridondanza in direzioni opposte. Nel complesso, Hasegawa e Miyata si sono occupati di amminoacidi, mentre Rumer di codoni. Combinati, questi approcci producono assegnazioni tra i codoni e i numeri di nucleoni degli amminoacidi, utili per la sistematizzazione. I codoni di stop non codificano per nessun amminoacido; pertanto, per includerli nella sistematizzazione, viene loro assegnato un numero di nucleoni pari a zero.

La chiave di attivazione. Tutti i modelli aritmetici considerati appaiono ulteriormente con la differenziazione tra blocchi e catene in tutti i 20 amminoacidi e con il successivo trasferimento di un nucleone dalla catena laterale al blocco nella prolina (Fig. 2b). La prolina è l'unica eccezione alla struttura generale degli amminoacidi: mantiene la sua catena laterale con due legami e ha un idrogeno in meno nel suo blocco. Il trasferimento citato nella prolina "uniforma" il numero di nucleoni del blocco a $73 + 1$ e riduce i nucleoni della catena a $42 - 1$. Di per sé, la distinzione tra blocchi e catene è puramente formale: non esiste una fase della sintesi proteica in cui le catene laterali degli amminoacidi si staccano dai blocchi standard. Pertanto, non c'è alcuna ragione naturale per il trasferimento di nucleoni nella prolina; può essere simulato solo nella mente di un ricevente per ottenere una serie di amminoacidi con struttura uniforme. Tale trasferimento di nucleoni appare quindi artificiale. Tuttavia, proprio questo sembra essere il suo scopo: proteggere i modelli da qualsiasi spiegazione naturale. Ridurre al minimo le possibilità di appellarsi all'origine naturale è una preoccupazione specifica in questo tipo di messaggistica, e questo problema sembra essere risolto perfettamente per il segnale nel codice genetico. Applicato sistematicamente, senza eccezioni, il trasferimento artificiale nella prolina consente un ordine olistico e aritmeticamente preciso nel codice. Agisce quindi come una "chiave di attivazione". Mentre la natura si occupa della prolina vera e propria che non produce il segnale nel codice, un destinatario intelligente trova facilmente la chiave e legge i messaggi in linguaggio aritmetico (vedi anche Discussione).

Decimalismo. Gli schemi aritmetici da descrivere sono validi in qualsiasi sistema numerico. Tuttavia, come si è visto, espressi nel sistema decimale posizionale, acquisiscono tutti una notazione distintiva. Pertanto, qui forniamo brevemente alcune informazioni rilevanti.

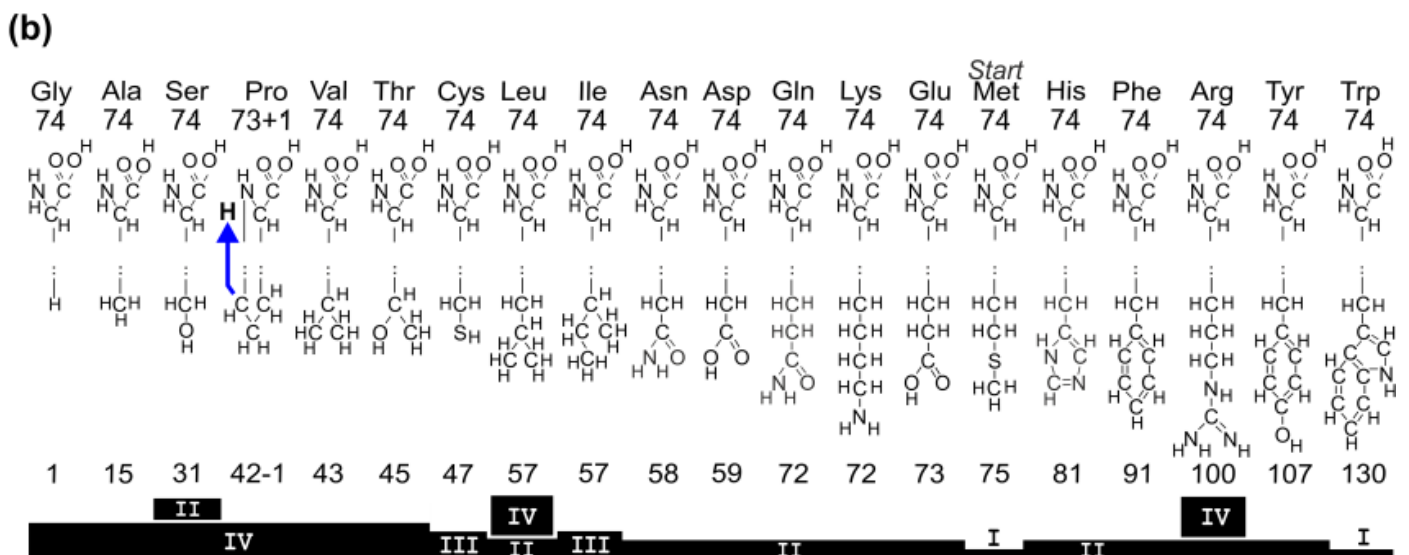
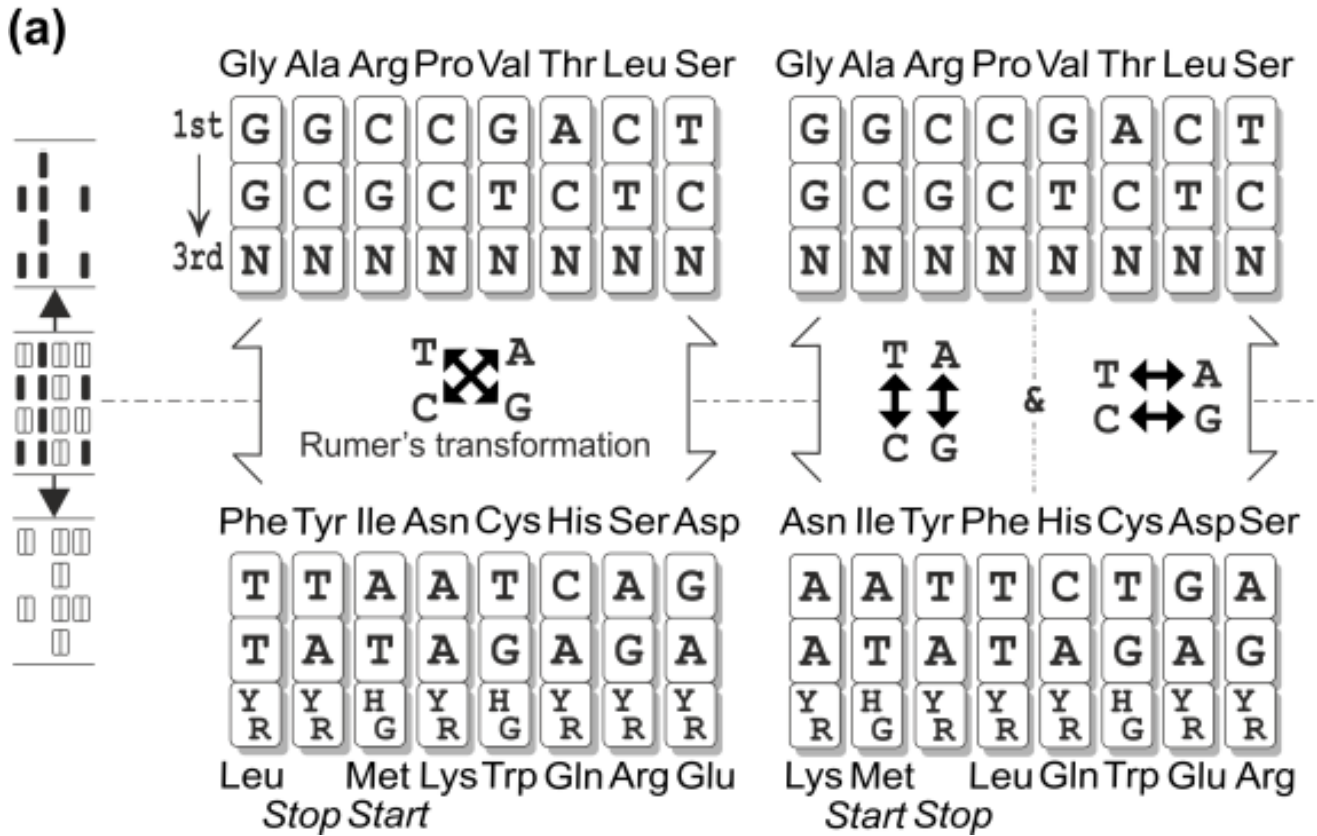


Figura 2. Osservazioni precedenti. (a) La bisezione di Rumer. Le famiglie intere vengono contrapposte a quelle divise, dividendo così il codice. I codoni delle famiglie opposte sono mappati l'uno sull'altro con l'insieme ordinato delle trasformazioni di Rumer e delle due semitrasformazioni. La trasformazione delle terze basi è banale in

La natura è indifferente ai linguaggi numerici escogitati dall'intelligenza per rappresentare le quantità, compreso lo zero. Un sistema numerico privilegiato è quindi un segno affidabile di artificialità. Intenzionalmente incorporato in un oggetto, un sistema privilegiato potrebbe dimostrarsi attraverso una notazione distintiva a qualsiasi destinatario che abbia a che fare con elementi enumerabili di quell'oggetto. Ad esempio, le simmetrie digitali dei numeri divisibili per il primo 037 esistono solo nel sistema decimale posizionale con concezione zero (Fig. 3). Così, i decimali distintivi 111, 222 e 333 appaiono ordinariamente 157, 336 e 515 nel sistema ottale. Questa caratteristica notarile fu segnalata da Pacioli (1508) poco dopo l'arrivo del sistema decimale in Europa. Un'analogia caratteristica a tre cifre esiste in alcuni altri sistemi, compreso quello quaternario (cfr. Appendice C).

Risultati

La struttura complessiva del segnale è illustrata nella Fig. 4, che può essere utilizzata come guida per ulteriori descrizioni. Il segnale è composto da schemi aritmetici e ideografici, dove le unità aritmetiche sono rappresentate dai nucleoni degli amminoacidi, mentre le basi dei codoni fungono da entità ideografiche. I modelli del segnale sono visualizzati in distinte disposizioni logiche del codice, aumentando così sia il contenuto informativo del segnale sia la sua significatività statistica. È interessante notare che tutti i modelli presentano lo stesso stile generale, come illustrato nella Fig. 4, con simboli identici in ogni componente del segnale (rappresentato da caselle). In particolare, disposizioni logiche distinte del codice e della chiave di attivazione producono uguaglianze esatte delle somme dei nucleoni, che presentano inoltre un decimalismo e sono accompagnate da trasformazioni di Rumer e/o mezze trasformazioni. Una di queste disposizioni porta inoltre a simmetrie ideografiche e semantiche. Tutti gli elementi del codice - 64 codoni, 20 amminoacidi, segni sintattici Start e Stop - sono coinvolti in ogni disposizione.

A differenza dei segnali radio, che si svolgono nel tempo e hanno quindi una struttura sequenziale, il segnale del codice genetico non ha un punto di ingresso, simile al messaggio pittorico delle placche dei pionieri (Sagan et al., 1972). Tuttavia, invece di fornire pittogrammi, il segnale del codice genetico fornisce schemi che non dipendono dai simboli visivi scelti per rappresentarli (siano essi simboli per le basi nucleotidiche o per la notazione di "numeri ostensivi"). Questi schemi costituiscono l'insieme organico, quindi non c'è un ordine unico nel presentarli. Inizieremo con la componente aritmetica per poi passare all'ideografia.

La componente aritmetica

Codice standard a grandezza naturale. Una disposizione logicamente semplice del codice è stata proposta da George Gamow nel suo tentativo di indovinare teoricamente le assegnazioni di codifica prima che il codice fosse decifrato in vitro (si veda Hayes, 1998). Uno dei suoi modelli, pur non prevedendo correttamente la mappatura effettiva, coincideva in modo notevole con una delle componenti del segnale. Gamow dispose i codoni in base alla loro composizione, poiché 20 combinazioni di quattro basi prese tre alla volta potevano dare origine a 20 amminoacidi (Gamow & Yčas, 1955). Inserendo in questa disposizione i numeri dei nucleoni, la chiave di attivazione e le poche condizioni di "congelamento", si scopre un bilanciamento totale dei nucleoni ornato da una sintassi decimale.

I codoni con basi identiche e uniche comprendono due insiemi più piccoli (Fig. 5a). Dimezzati, entrambi gli insiemi mostrano l'equilibrio delle catene laterali con $703 = 037 \times 019$ nucleoni in ciascuna metà e l'equilibrio delle molecole intere con $1665 = 666 + 999 \times 1$ nucleoni. È importante notare che il dimezzamento non è arbitrario. I codoni sono opposti dalla trasformazione di Rumer insieme alla semitrasformazione TC, AG nel primo insieme e TA, CG nel secondo insieme.

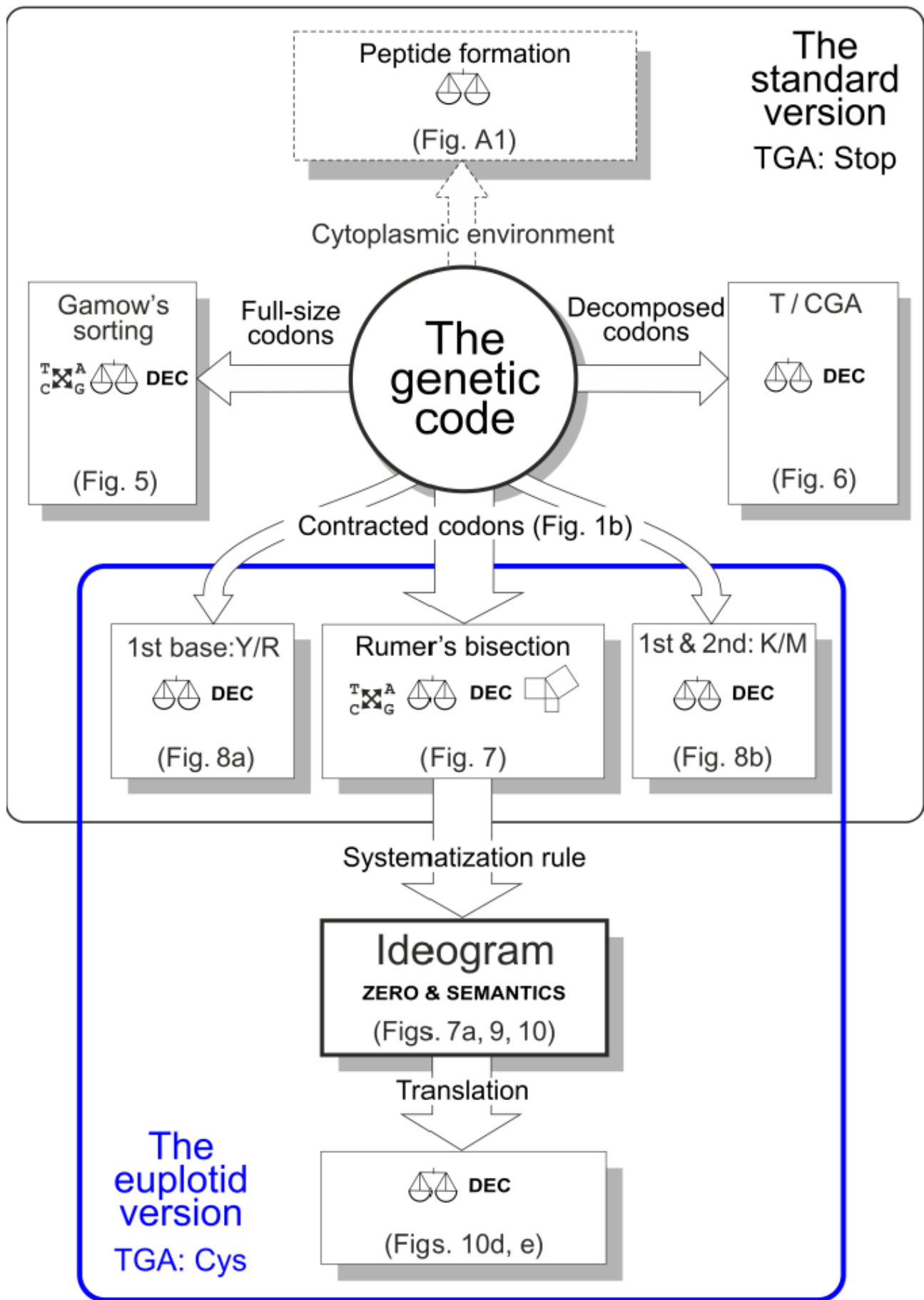
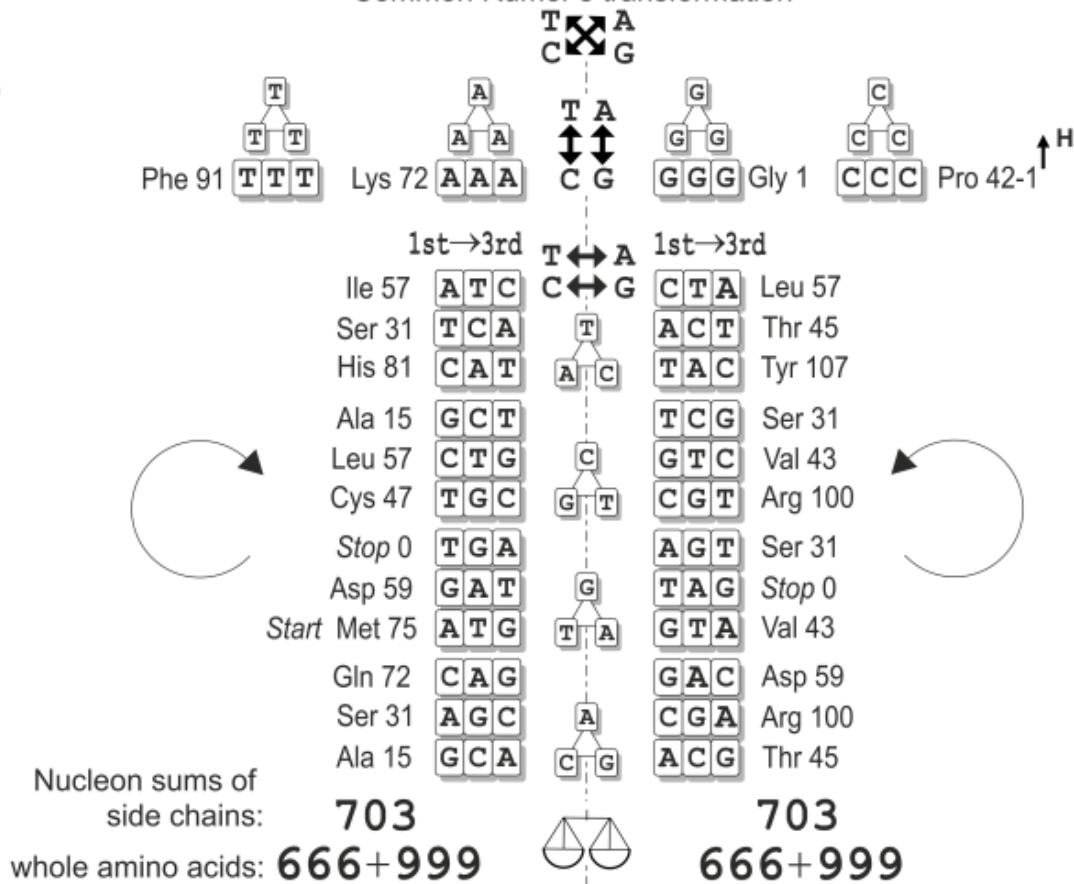


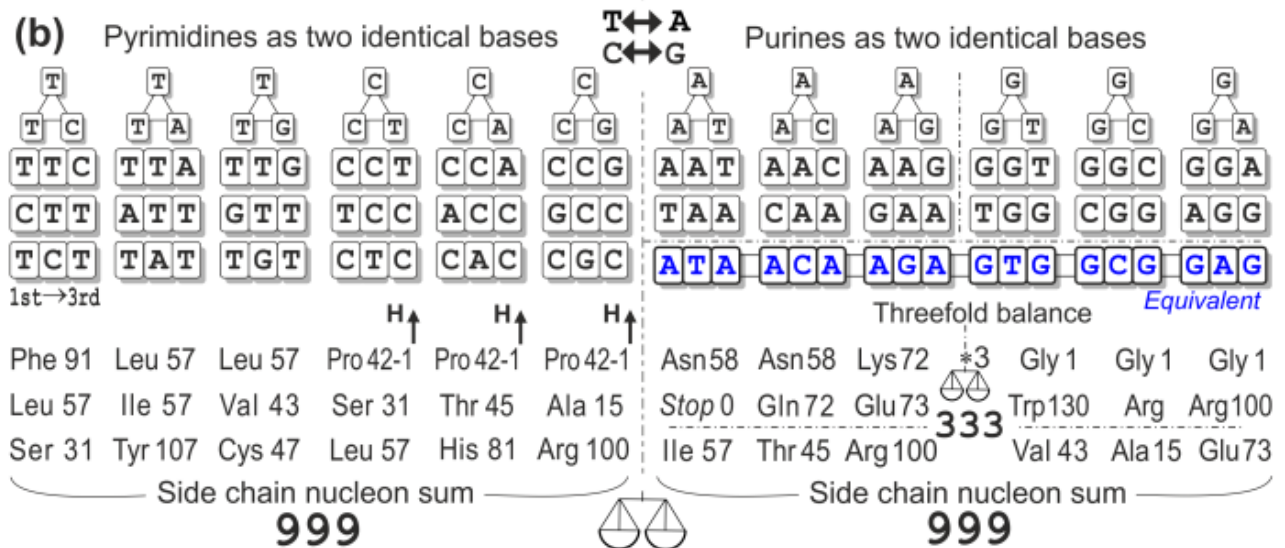
Fig. 4. Struttura del segnale. Tutti i dettagli sono discussi in sequenza nel testo. L'immagine delle scale rappresenta precise uguaglianze di nucleoni. DEC indica la notazione decimale distintiva delle somme di nucleoni. Il riquadro tratteggiato indica l'equilibrio citoplasmatico (vedi Appendice D), l'unico modello mantenuto dalla prolina e dall'ambiente cellulare. Tutti gli altri schemi sono abilitati dalla "chiave di attivazione" e sono validi per gli amminoacidi liberi. K sta per {T, G}, M sta per {A, C}. Sebbene tutti e tre i tipi di trasformazione agiscano nei modelli, per semplicità viene indicata solo la trasformazione di Rumer.

Common Rumer's transformation

(a)



(b)



(c) Pyrimidines as unique bases

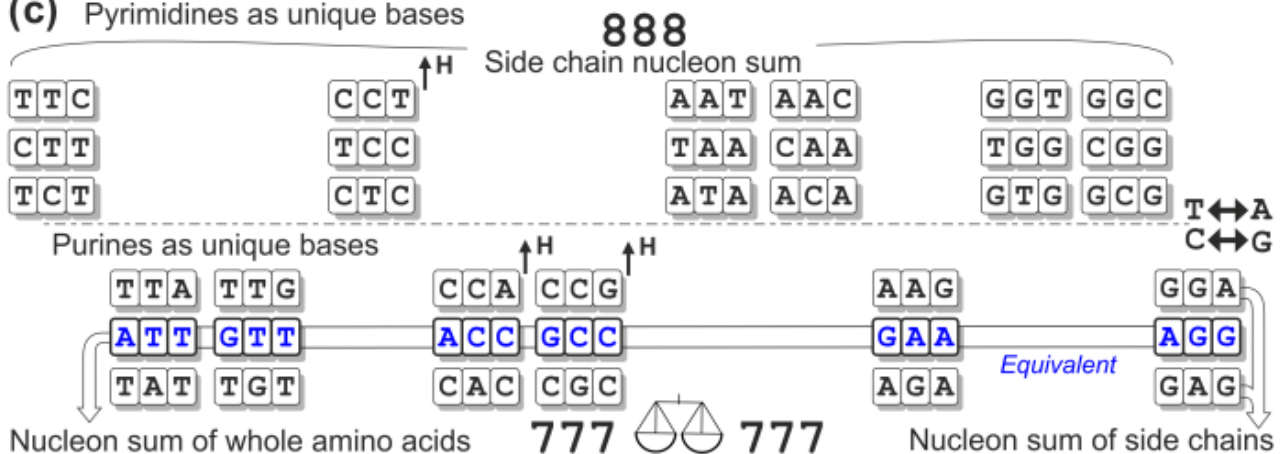


Fig. 5. Ordinamento di Gamow dei codoni in base alla composizione delle basi nucleotidiche. Le combinazioni di basi (indicate su cornici triangolari) producono tre insiemi: 4 codoni con tre basi identiche, 24 codoni con basi uniche e 36 codoni con due basi identiche. (a) Il primo e il secondo insieme dimezzati dall'asse verticale con le trasformazioni di Rumer e dimezzamento e la trasformazione Spin Antispin indicata con frecce circolari. Applicate alle cornici triangolari, queste frecce definiscono la sequenza di basi nei codoni. Si noti che mentre qualsiasi somma di blocchi (con la chiave di attivazione applicata) è divisibile per 037 poiché ogni blocco ha $74 = 2 \times 037$ nucleoni, le somme di catena non sono limitate in questo modo. (b) Il terzo insieme dimezzato a seconda che le basi identiche siano purine o pirimidine. (c) Il terzo insieme dimezzato con asse orizzontale a seconda che le basi uniche siano purine o pirimidine.

La trasformazione Spin \rightarrow Antispin non influisce sul primo insieme, ma congela gli elementi del secondo. Rimane un solo grado di libertà, poiché non esistono trasformazioni reversibili che possano collegare i due insiemi, quindi uno di essi è libero di scambiarsi intorno all'asse. L'equilibrio appare in uno dei due stati alternativi.

Il terzo insieme comprende i codoni con due basi identiche. Se dimezzato a seconda che si tratti di purine o pirimidine, indipendentemente dal tipo di base unica, questo insieme mostra l'equilibrio $999 = 999$ di catene laterali (Fig. 5b). Inoltre, tale dimezzamento mantiene la semitrasformazione di Rumer e una delle semitrasformazioni. A sua volta, la metà destra dell'insieme è triplicemente bilanciata. I codoni con adenina affiancata, guanina affiancata e codoni palindromici costituiscono tre parti uguali con 333 nucleoni ciascuna.

Nella Fig. 5c lo stesso insieme viene dimezzato a seconda che le basi uniche siano purine o pirimidine, questa volta indipendentemente dal tipo di basi identiche. Anche se non bilanciate, queste metà mostrano ancora una volta una sintassi decimale distintiva con 888 e $1110 = 111 + 999 \times 1$ nucleoni. Il decimalismo di una di queste somme è algebricamente dipendente, poiché nel caso precedente (Fig. 5b) è noto che la somma dell'intero insieme è divisibile per 037; se una parte di questo insieme è decimale distintiva, l'altra lo sarà automaticamente. Tuttavia, in questo caso si nota un modello indipendente. In particolare, una parte del precedente triplice equilibrio ha un equivalente in una metà, dove gli stessi amminoacidi sono rappresentati da codoni sinonimi (Fig. 5b e c). Intere molecole di questo equivalente - 333 nucleoni di catena laterale e 444 di blocco standard - sono bilanciate con 777 nucleoni di catena nel resto del sottoinsieme.

Si noti che tutte queste notazioni distintive delle somme di nucleoni appaiono solo nel sistema decimale posizionale. La notazione posizionale è così abituale nella nostra cultura che la maggior parte dei suoi utenti difficilmente ricorda la regola abbastanza complessa che sta alla base della codifica dei numeri come $a_{n-1} \times q^{n-1} + \dots + a_1 \times q^1 + a_0 \times q^0$, dove $q = 10$ nel caso del sistema decimale, n è la quantità di cifre nella notazione e a_i - le cifre 0-9 che rimangono nella notazione finale.

Codice standard decomposto. Un'altra disposizione del codice è data dalla decomposizione dei 64 codoni a grandezza naturale. In questo modo si ottengono 192 basi separate e si rivela un modello dello stesso tipo di quello a grandezza naturale. Le basi identiche formano quattro gruppi di 48 basi ciascuno. Ogni base conserva l'amminoacido o lo stop del suo codone originale (Fig. 6a). In questo modo, i quattro insiemi ottengono le loro somme individuali di catene e blocchi di nucleoni.

In totale, ci sono $222 + 999 \times 10$ nucleoni di catena laterale nel codice decomposto - ovviamente, il triplo della somma totale nel caso precedente a grandezza naturale (con la chiave di attivazione ancora applicata). Solo una combinazione dei quattro insiemi mostra un decimalismo distintivo delle somme dei nucleoni della catena laterale. Si tratta di $666 + 999 \times 2$ nucleoni nell'insieme T e di $555 + 999 \times 7$ nucleoni nell'insieme CGA (Fig. 6b). Nel frattempo, ci sono esattamente $222 + 999 \times 10$ nucleoni di blocco nell'insieme CGA (si noti che gli insiemi hanno somme di blocco disuguali a causa del diverso accumulo di stop). Pertanto, sebbene i nucleoni a catena siano più numerosi dei nucleoni a blocco nell'insieme del codice, essi sono perfettamente bilanciati con la loro parte CGA.

Codice contratto e regola di sistematizzazione. In un certo senso, la contrazione della serie di codoni (vedi Fig. 1b) è un'operazione logicamente opposta alla decomposizione. Oltre a mostrare nuovi schemi aritmetici, il codice contratto rivela anche la componente ideografica del segnale. La regola di sistematizzazione che porta all'ideografia combina i risultati di Rumer (1966) e di Hasegawa & Miyata (1980) ed è di natura simmetrica (shCherbak, 1993). Le serie contratte sono ordinate in quattro gruppi in base alla loro ridondanza; all'interno di questi gruppi sono allineate una accanto all'altra in ordine di variazione monotona (ad esempio, crescente) del numero di nucleoni. Le serie stesse sono poi disposte in modo antisimmetrico (ad esempio, in ordine di numero di ridondanza decrescente). La serie di stop viene posta all'inizio del suo insieme, rappresentando lo zero nella sua posizione speciale. Infine, la bisezione di Rumer oppone l'insieme IV agli insiemi III, II e I. La disposizione risultante è mostrata nella Fig. 7 per il codice euplotidico, con l'ideografia delle basi dei codoni

(vedi sezione successiva) nella Fig. 7a e gli schemi aritmetici degli amminoacidi (condivisi da entrambe le versioni del codice) nella Fig. 7b.

Un nuovo equilibrio si trova nell'insieme congiunto III, II, I. I nucleoni della catena laterale di tutti gli amminoacidi sono equiparati ai loro blocchi standard: $111 + 999 \times 1 = 111 + 999 \times 1$ (Fig. 7b). Questo schema si manifesta come l'anticorrelazione menzionata da Hasegawa & Miyata (1980). La somma dei nucleoni della catena di tutte le serie del codice è inferiore alla somma di tutti i blocchi. Solo un sottoinsieme di serie che codifica principalmente amminoacidi più grandi può pareggiare i propri blocchi. Ciò accade esattamente nella serie congiunta III, II, I. Di conseguenza, gli amminoacidi più piccoli vengono lasciati nell'insieme di ridondanza IV.

Nel frattempo, ci sono 333 nucleoni di catena e 592 di blocco e $333 + 592 = 925$ nucleoni di molecole intere nell'insieme IV. Con la cancellazione di 037, si ottiene $3^2 + 4^2 = 5^2$ - rappresentazione numerica del triangolo egizio, forse come simbolo dello spazio bidimensionale. Per inciso, le serie di codoni nell'ideogramma (Fig. 7a) sono disposte nel piano piuttosto che in modo lineare come nella genomica.

La bisezione di Rumer si basa sulla ridondanza e fa quindi uso di terze posizioni nelle serie di codoni. Anche le divisioni del codice contratto in base alla prima posizione e alla posizione centrale rivelano schemi simili (Fig. 8). Un altro fenomeno aritmetico presumibilmente legato al segnale - l'equilibrio citoplasmatico - è descritto nell'Appendice D.

Pertanto, il codice standard rivela modelli dello stesso stile, ma algebricamente indipendenti, simultaneamente nelle rappresentazioni scomposte, a grandezza naturale e contratte (cfr. Fig. 4). È un compito algebrico non banale trovare una soluzione che mappi gli amminoacidi e i segni sintattici ai codoni in modo simile. Normalmente ciò richiederebbe una notevole potenza di calcolo.

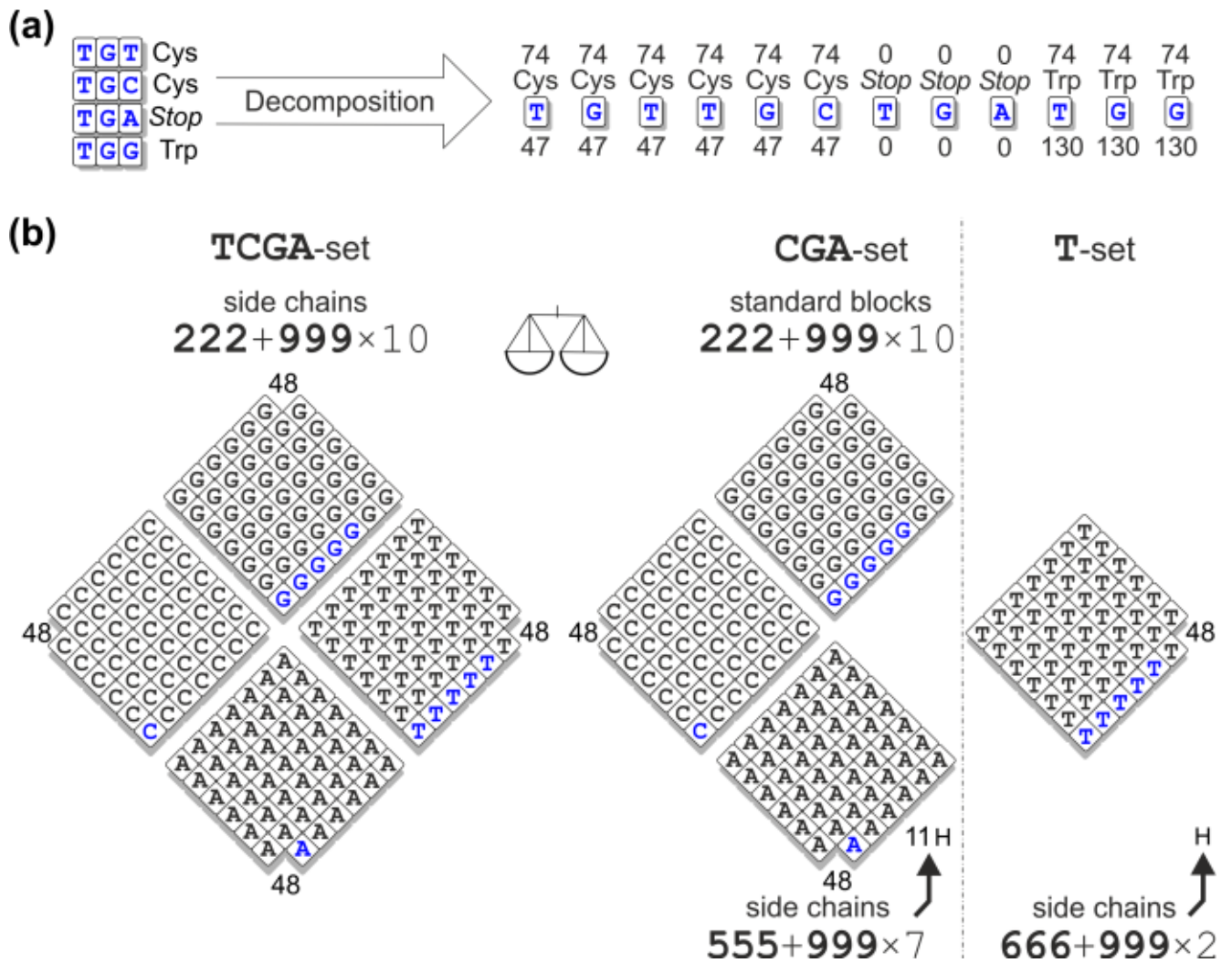
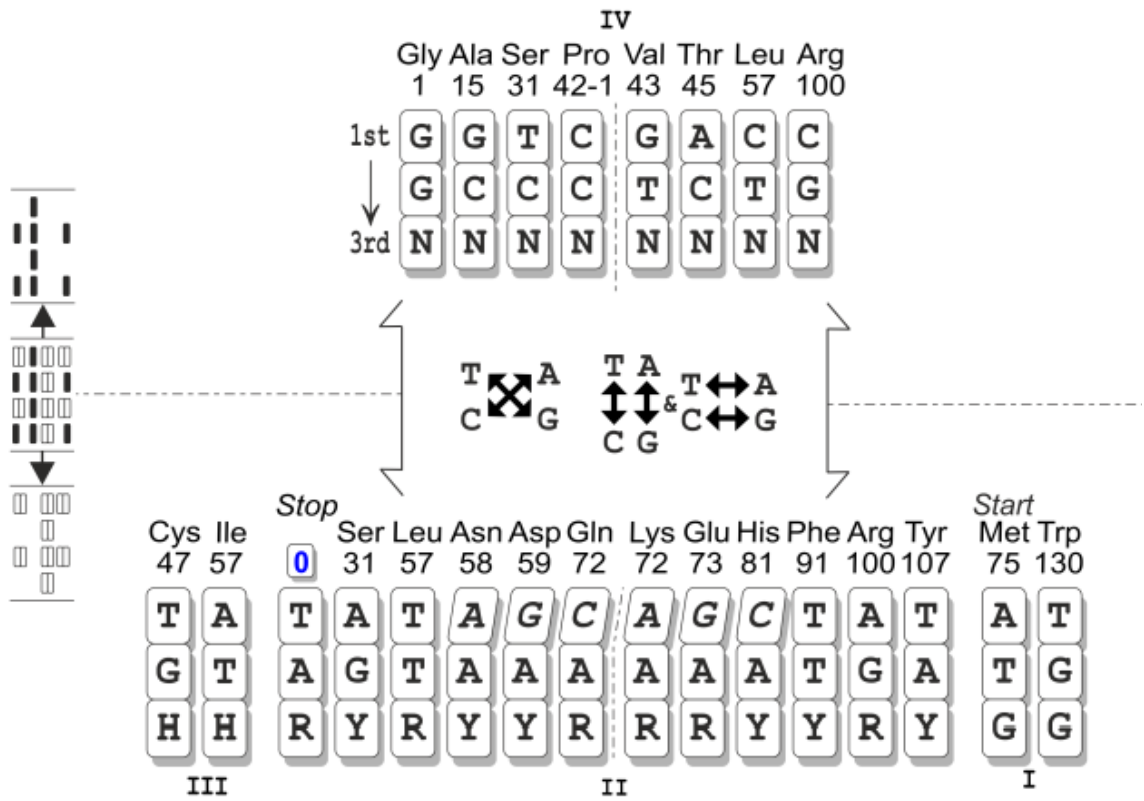


Fig. 6. Il codice standard decomposto. (a) Decomposizione mostrata per una famiglia di codoni. Tre basi T contribuiscono con tre molecole di Cys all'insieme T; una base A contribuisce con uno Stop all'insieme A e così via per l'intero codice. (b) Le basi identiche sono ordinate in quattro insiemi, indipendentemente dalla loro posizione nei codoni. Gli insiemi sono mostrati due volte per comodità.

(a)

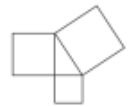
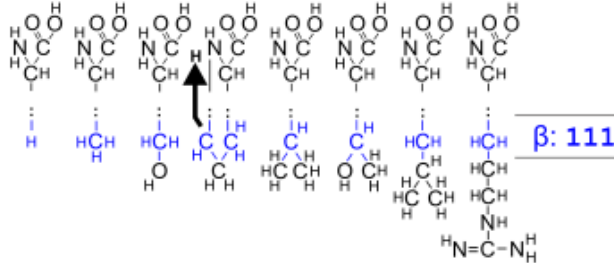


(b) Whole molecules: 75 89 105 115 117 119 131 174 = 925

$25=5^2$

blocks: Gly 74 Ala 74 Ser 74 Pro 73+1 Val 74 Thr 74 Leu 74 Arg 74 = 592

$16=4^2$



chains: 1 15 31 42-1 43 45 57 100 = 333

$9=3^2$

IV

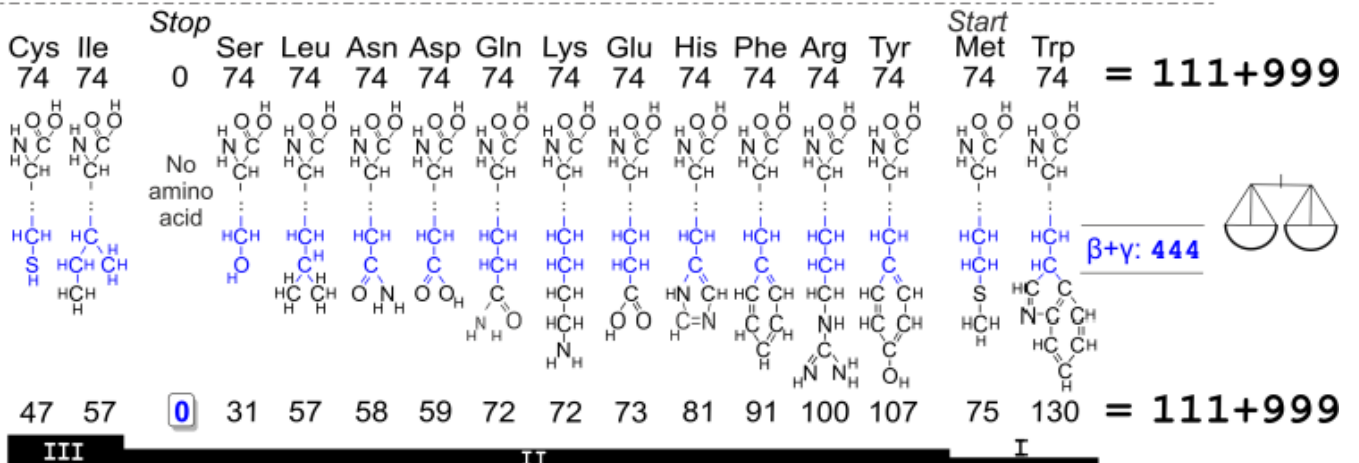


Fig. 7. Il codice euplotidico contratto con l'applicazione della regola di sistematizzazione (confronto con la Fig. 2). (a) La disposizione risultante delle serie di codoni contratti che formano l'ideogramma. L'allineamento laterale delle serie verticali produce tre stringhe orizzontali di basi posizionate alla pari. Gln e Lys hanno lo stesso numero di nucleoni; l'ambiguità nel loro posizionamento è eliminata dalle simmetrie considerate in seguito. (b) Lo sfondo aritmetico dell'ideogramma (valido anche per la versione standard, in quanto contribuisce con un altro zero alla serie III, II, I). Per i livelli di catena laterale e si veda la **Discussione**.

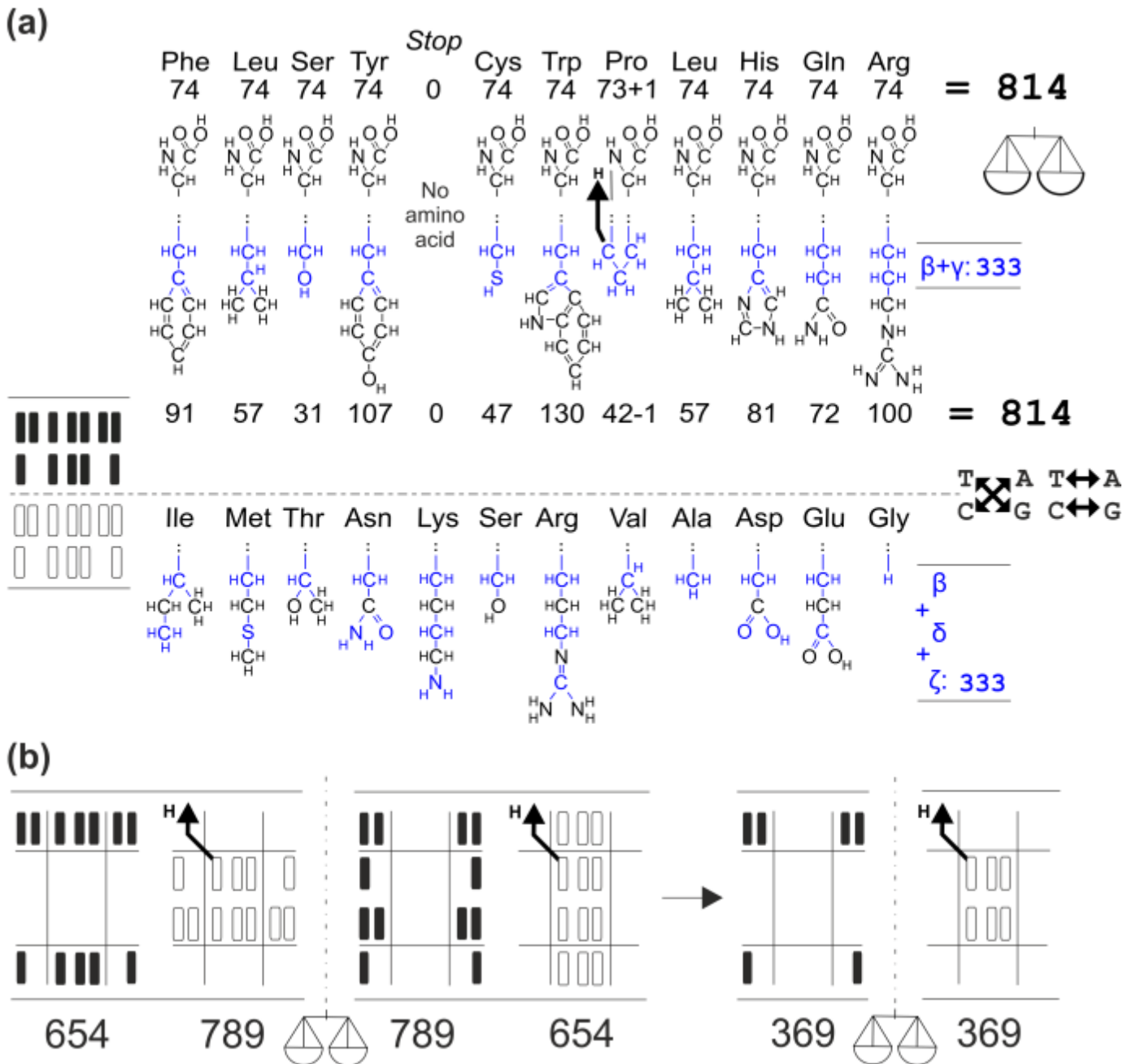


Fig. 8. Schemi aritmetici aggiuntivi del codice contratto (condivisi da entrambe le versioni del codice). (a) Il codice è diviso a seconda che le prime basi siano purine o pirimidine. Si ottengono così due serie con un numero uguale di serie. La metà con le pirimidine in prima posizione rivela un nuovo equilibrio di catene e blocchi analogo a quello della Fig. 7b. Un'altra metà dipende algebricamente solo dalla somma decimale dei suoi livelli β , δ , ζ , vedi Discussione. (b) Il codice è diviso a seconda che le prime basi siano K o M (sinistra) o che le basi centrali siano K o M (centro). Entrambe le divisioni producono metà con somme di nucleoni a catena identiche. Come conseguenza algebrica di queste divisioni, le serie con K in prima posizione e in posizione centrale e le serie con M in prima posizione e in posizione centrale sono bilanciate a catena (a destra). Ciascuna delle tre divisioni è accompagnata da

semitrasformazioni e, cosa notevole, produce anche un numero uguale di serie in ciascuna metà. Questo schema è l'unico che non presenta la divisibilità per 037. Tuttavia, tutti e tre i numeri - 654, 789 e 369 - sono di nuovo specifici in notazione decimale, dove le cifre in ciascuno di essi appaiono come progressioni aritmetiche.

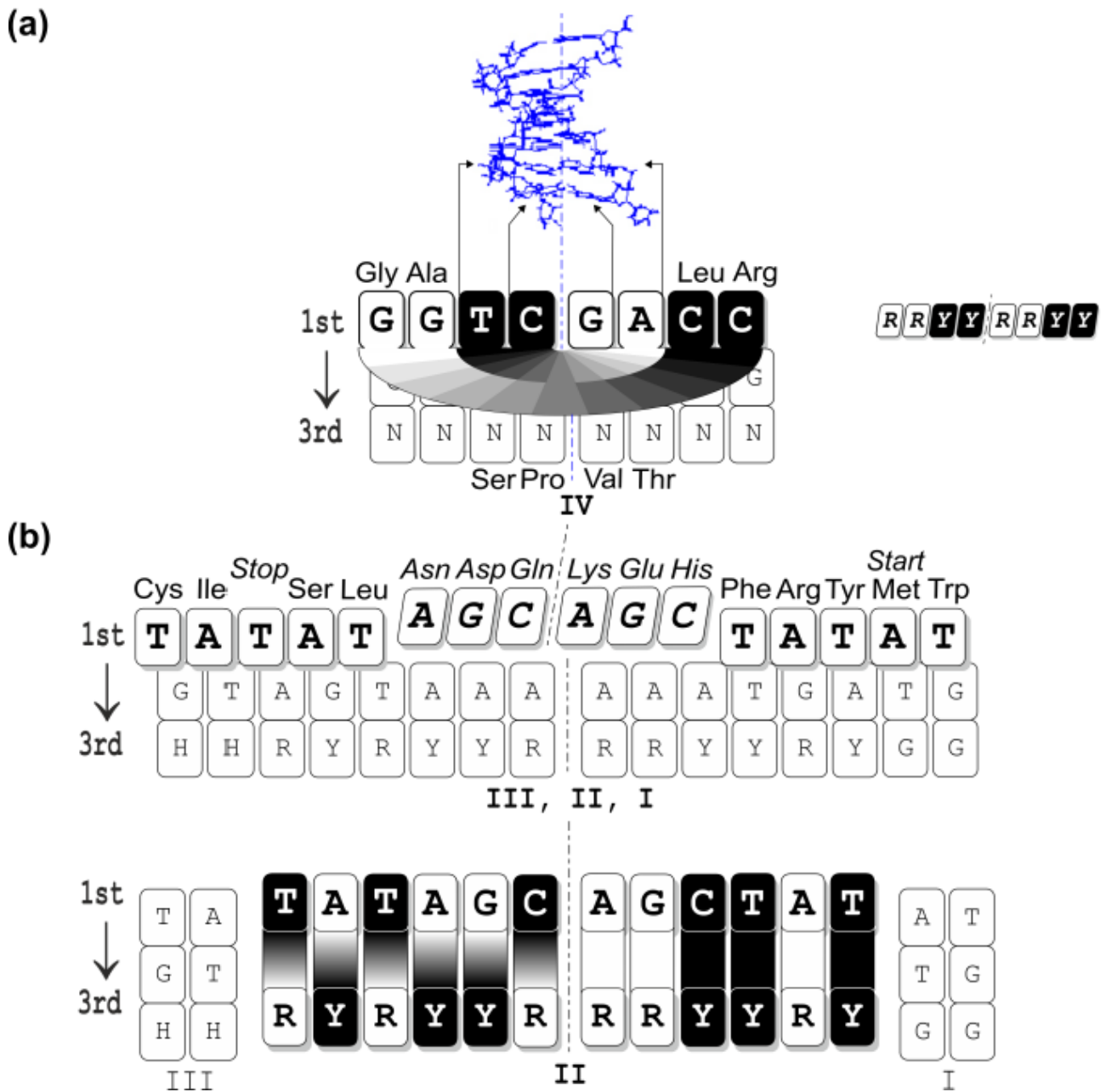


Fig. 9. Schemi delle stringhe superiori corte (a) e lunghe (b). Le stringhe sono disposte con la stessa serie di simmetrie: simmetria speculare (indicata con l'asse verticale centrale), simmetria di traslazione (indicata con lettere in corsivo e cornici oblique) e inversione purina-pirimidina (indicata con una sfumatura di colore, dove il nero e il bianco rappresentano rispettivamente le pirimidine e le purine). L'immagine del DNA in alto illustra la possibile interpretazione della stringa corta (vedi Discussione).

La componente ideografica

Stringhe superiori. Ci riferiamo al prodotto della sistematizzazione nella Fig. 7a come ideogramma. L'ideogramma del codice genetico si basa sulle simmetrie delle sue stringhe (shCherbak, 1988). Le stringhe vengono lette in serie contratte.

La stringa corta superiore presenta simmetrie speculari, di traslazione e di inversione (Fig. 9a). Le sue basi sono invarianti sotto l'operazione combinata della simmetria speculare e dell'inversione della base complementare del tipo. Un modello minimo della simmetria di traslazione è rappresentato dal quadruplo RRYY.

Le stesse tre simmetrie dispongono la lunga stringa superiore (Fig. 9b). La coppia di sequenze TATAT affiancate è simmetrica a specchio. La coppia di codoni centrali AGC forma un modello minimo della simmetria di traduzione. La prima e la terza base dell'insieme di ridondanza II sono interconnesse in modo assialsimmetrico con l'inversione purina-pirimidina e la sua operazione opposta - la trasformazione unitaria che non produce alcuno scambio.

Stringhe centrali. Collocate coassialmente, le stringhe centrali corte e lunghe appaiono interconnesse con l'inversione purina-pirimidina (Fig. 10a). Entrambe le stringhe presentano una simmetria speculare purina-pirimidina. La stringa lunga mantiene la simmetria speculare anche per le basi ordinarie.

I codoni della stringa corta CCC e TCT rompono la simmetria speculare delle basi ordinarie, ma condividono una caratteristica palindroma, ossia l'invarianza della direzione di lettura. Questa caratteristica ripristina la simmetria speculare, questa volta di tipo semantico (Fig. 10b). Come nel caso precedente, ci si aspetta che due stringhe centrali condividano lo stesso insieme di simmetrie. Pertanto, la simmetria semantica dei codoni palindromici affiancati da basi G può indicare una caratteristica simile nella stringa lunga. In effetti, la simmetria semantica si riscontra nella cornice di lettura della tripletta che inizia dopo la base G che la affianca (Fig. 10c). Questa cornice di lettura è notevole per la disposizione regolare di tutti i segni sintattici del codice euplotidico - entrambi i codoni Stop e il codone Start ripetuti due volte. La cornice di lettura mostra la simmetria semantica speculare degli antonimi con il codone AAA omogeneo al centro.

I codoni di questa cornice di lettura sono simboli puramente astratti, dato che vengono letti in serie contratte. Tuttavia, sono regolarmente incrociati con gli stessi codoni nell'ideogramma, rafforzando così la simmetria semantica e rendendo unica la cornice attuale (Fig. 10c). Inoltre, la direzione di lettura ora si distingue, poiché tale "cruciverba" scompare se letto in senso opposto, sebbene il palindromo stesso rimanga invariato.

È interessante notare che la stringa della tripletta nella Fig. 10c è scritta con i simboli del codice all'interno del codice stesso. Ciò implica che la mappatura signalharboring doveva essere proiettata in via preliminare (vedi Discussione). Inoltre, la traduzione di questa stringa con il codice stesso rivela l'equilibrio $222 = 222$ di catene e blocchi (Fig. 10d). Un ulteriore palindromo nella cornice spostata di una posizione (Fig. 10e) riproduce la somma delle catene di 222, confermando che l'ideogramma è correttamente "sintonizzato" sulla versione euplotidica: TGA sta qui per Cys, non per Stop del codice standard.

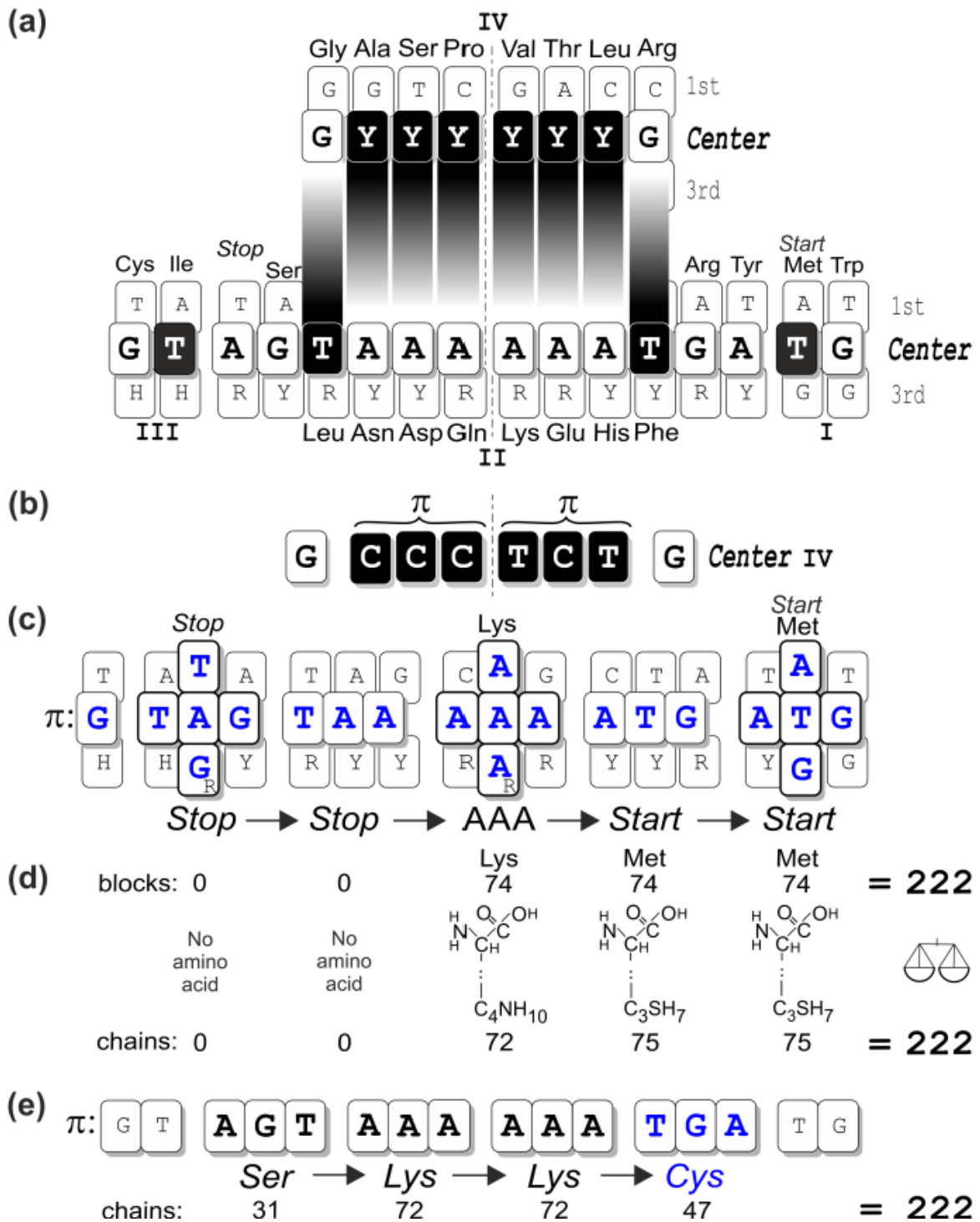


Fig. 10. Schemi delle stringhe centrali corte (a, b) e lunghe (a, c, d, e). Entrambe le stringhe sono disposte con simmetria speculare purina-pirimidina, inversione purina-pirimidina e simmetria semantica. Le prime due sono indicate come nella Fig. 9, che denota il palindromo.

Discussione

Artificialità. Per essere considerati inequivocabilmente un segnale intelligente, i modelli del codice devono soddisfare i due criteri seguenti: (1) devono essere altamente significativi dal punto di vista statistico e (2) non solo devono possedere caratteristiche simili a quelle dell'intelligenza (Elliott, 2010), ma devono essere incoerenti in linea di principio con qualsiasi processo naturale, sia esso darwiniano (Freeland, 2002) o lamarckiano (Vetsigian et al, 2006), guidato dalla biosintesi degli aminoacidi (Wong, 2005), dai cambiamenti genomici (Sella & Ardell, 2006), dalle affinità tra (anti)codoni e aminoacidi (Yarus et al., 2009), dalla selezione per l'aumento della diversità delle proteine (Higgs, 2009), dall'energetica delle interazioni codone-anticodone (Klump, 2006; Travers, 2006) o da vari meccanismi pre-traslazionali (Wolf & Koonin, 2007; Rodin et al., 2011).

Il test statistico per il primo criterio è descritto nell'Appendice B e mostra che i modelli descritti sono altamente significativi. Il secondo criterio potrebbe sembrare non verificabile, in quanto i pattern potrebbero derivare da un processo naturale attualmente sconosciuto. Ma questo criterio equivale a chiedersi se sia possibile incorporare nel codice schemi informativi tali da poter essere interpretati inequivocabilmente come una firma intelligente. La risposta sembra essere affermativa e un modo per farlo è rendere i modelli virtuali, non reali. Questo è esattamente ciò che si osserva nel codice genetico. Gli equilibri rigidi e la loro sintassi decimale appaiono solo con l'applicazione della "chiave di attivazione". Fisicamente, non ci sono equilibri rigidi nel codice (ad esempio, nella Fig. 5b si avrebbe $1002 \neq 999$ invece di $999 = 999$). Il trasferimento artificiale di un nucleone nella prolina attiva gli schemi aritmetici e li rende quindi virtuali. Questo è anche il motivo per cui interpretiamo la notazione distintiva come un'indicazione di decimalismo, piuttosto che come un requisito fisico (ancora sconosciuto) per cui le somme dei nucleoni devono essere multipli di 037: in generale, fisicamente non esiste tale molteplicità nel codice. A sua volta, il sistema numerico preferito dal punto di vista notarile è di per sé un forte segno di artificiosità. Vale anche la pena notare che tutti i decimali a tre cifre - 111, 222, 333, 444, 555, 666, 777, 888, 999 (oltre allo zero, vedi sotto) - sono rappresentati almeno una volta nel segnale, il che sembra anch'esso una caratteristica intenzionale.

Tuttavia, si potrebbe ipotizzare che la massa degli amminoacidi sia guidata dalla selezione (o da qualsiasi altro processo naturale) per essere distribuita nel codice in un modo particolare che porta a un'uguaglianza di massa approssimativa, rendendo così il rigoroso bilanciamento dei nucleoni solo un probabile epifenomeno. Ma è difficilmente immaginabile come un processo naturale possa guidare la distribuzione

della massa in rappresentazioni astratte del codice in cui i codoni sono scomposti in basi o contratti dalla ridondanza. Inoltre, le uguaglianze tra i nucleoni valgono per gli amminoacidi liberi, eppure in queste molecole libere le catene laterali e i blocchi standard dovevano essere trattati da questo processo separatamente. Inoltre, nessun processo naturale può guidare la distribuzione della massa per produrre l'equilibrio della Fig. 10d: gli amminoacidi e i segni sintattici che compongono questo equilibrio sono completamente astratti, poiché sono prodotti dalla traduzione di una stringa letta attraverso i codoni.

Un altro modo per rendere i modelli irriducibili agli eventi naturali è quello di coinvolgere la semantica, poiché nessun processo naturale è in grado di interpretare simboli astratti. Va notato che le nozioni di simboli e significati sono talvolta utilizzate in senso naturale (Eigen & Winkler, 1983), soprattutto nel contesto della biosemiotica (Barbieri, 2008) e dei codici molecolari (Tlusty, 2010). Il codice genetico stesso è considerato una "convenzione naturale" che mette in relazione i simboli (codoni) con i loro significati (aminoacidi). Tuttavia, questi approcci fanno una distinzione tra la semantica organica dei codici molecolari e la semantica interpretativa o linguistica propria dell'intelligenza (Barbieri, 2008). Proprio quest'ultimo tipo di semantica si rivela nel segnale del codice genetico. Esso si manifesta non solo nella simmetria dei segni sintattici antonimi (Fig. 10c), ma anche nel simbolo dello zero. Per il macchinario molecolare genetico non esiste lo zero, esistono triplette nucleotidiche riconosciute stericamente da fattori di rilascio nel ribosoma. Lo zero - suprema astrazione dell'aritmetica - è il significato interpretativo assegnato agli Stop-codon, e la sua correttezza è confermata dal fatto che, collocato nella sua giusta posizione frontale, lo zero mantiene tutte le simmetrie dell'ideogramma. Quindi, sommatore banale negli equilibri, lo zero appare però come numero ordinale nell'ideogramma. In altre parole, oltre a essere parte integrante del sistema decimale, lo zero agisce anche come simbolo individuale nel codice.

In totale, non solo il segnale in sé rivela caratteristiche simili a quelle dell'intelligenza - rigorose uguaglianze tra nucleoni, la loro distintiva notazione decimale, le trasformazioni logiche che accompagnano le uguaglianze, il simbolo dello zero e le simmetrie semantiche, ma il metodo stesso della sua estrazione comporta operazioni astratte - la considerazione di molecole idealizzate (libere e non modificate), la distinzione tra i loro blocchi e le loro catene, la chiave di attivazione, la contrazione e la scomposizione dei codoni. Nel complesso, tutti questi aspetti indicano la natura artificiale dei modelli.

Sebbene il sistema decimale nel segnale possa sembrare una coincidenza serendipica, ci sono poche spiegazioni possibili, dall'anatomia a 10 cifre come quasi-ottimale evolutivo per gli esseri bilaterali (Dennett, 1996) al fatto che ci sono convenientemente

$74 = 2 \times 037$ nucleoni nei blocchi standard di α aminoacidi. Inoltre, il sistema decimale condivide la simmetria digitale a triplette con quello quaternario (vedi Appendice C), stabilendo un legame con il linguaggio "nativo" del DNA. Del resto, anche alcuni dei messaggi inviati dalla Terra includevano il sistema decimale (Sagan et al., 1978; Dumas & Dutil, 2004), anche se non si supposeva che dovessero essere ricevuti necessariamente da extraterrestri a 10 cifre. Qualunque sia la ragione effettiva del sistema decimale nel codice, sembra che sia stato inventato al di fuori del Sistema Solare già diversi miliardi di anni fa.

Due versioni del codice. La versione del codice quasi simmetrica con schemi aritmetici funge da codice standard universale. Con questo codice a portata di mano è intuitivamente facile dedurre la versione simmetrica con la sua ideografia. Viceversa, se la versione simmetrica fosse quella universale, difficilmente sarebbe possibile dedurre il codice quasi simmetrico con tutti i suoi schemi aritmetici. Pertanto, con la sola versione standard è possibile "ricevere" entrambe le componenti aritmetiche e ideografiche del segnale, anche se la versione simmetrica non è stata trovata in natura. Ci sono due possibili ragioni per cui si trova effettivamente nei ciliati euplotidi: o in origine, quando la Terra è stata seminata, c'erano entrambe le versioni del codice e una di esse è rimasta attualmente nei ciliati euplotidi, oppure in origine c'era solo la versione standard e in seguito una modifica casuale nella discendenza degli euplotidi ha coinciso con la versione simmetrica.

Per quanto riguarda le altre rare versioni conosciute del codice, esse non sembrano avere un profondo insieme di modelli, né essere facilmente deducibili dal codice standard. Come comunemente accettato, rappresentano deviazioni casuali successive del codice standard causate da intermedi ambigui o da catture di codoni (Moura et al., 2010).

Incorporazione del segnale. Per ottenere un codice con una firma si potrebbero cercare tutte le varianti di mappatura e selezionare quella "più interessante". Tuttavia, questo metodo non è pratico (almeno con le attuali strutture di calcolo terrestri), dato il numero astronomicamente enorme di codici varianti. In un'alternativa più realistica, l'insieme dei pattern del segnale viene proiettato preliminarmente come un sistema di espressioni algebriche che viene poi risolto con relativa facilità per dedurre la mappatura del codice. Pertanto, tutti i pattern descritti potrebbero essere rappresentati post factum come un sistema di espressioni diofantine (cioè equazioni e disuguaglianze che ammettono solo soluzioni intere), e l'analisi di questo sistema mostra che determina in modo univoco la mappatura tra le serie di codoni e i numeri di nucleoni, compresi gli zeri per i codoni di stop (vedi Appendice E). Sebbene alcuni

amminoacidi abbiano numeri di nucleoni uguali, come nel caso di Leu e Ile, o Lys e Gln, anche questi non sono intercambiabili, come suggerisce la notazione distintiva delle somme dei nucleoni in β , γ e altri livelli posizionali delle catene laterali nel codice contratto (Figg. 7b e 8a). Anche in questo caso vale la chiave di attivazione (si noti che β - e δ -carboni nella prolina sono posizionalmente equivalenti). La nomenclatura chimica standard degli atomi di carbonio viene qui estesa per indicare le posizioni di altri atomi nodali. Il decimalismo in diverse combinazioni di livelli aggira la dipendenza algebrica e utilizza la struttura chimica degli amminoacidi in modo più efficiente.

Questi schemi all'interno delle catene laterali vanno ancora più in profondità nella struttura chimica. Alcuni degli amminoacidi canonici - His, Arg e Trp - potrebbero esistere in forme tautomeriche neutre alternative che differiscono per la posizione di un atomo di idrogeno nelle loro catene laterali (Taniguchi & Hino, 1981; Rak et al., 2001; Li & Hong, 2011). Sebbene alcuni di questi tautomeri si verificano molto raramente a pH citoplasmatico (come nel caso del tautomero indolenina di Trp mostrato in Fig. 7b), tutti i tautomeri neutri sono legittimi se si considerano molecole libere idealizzate, e prenderne solo uno introdurrebbe un'arbitrarietà. Si noti, tuttavia, che mentre un tautomero di Trp mantiene i modelli nella Fig. 7b, un altro fa il lavoro nella Fig. 8a, mentre qualsiasi tautomero neutro di His e Arg potrebbe essere preso in entrambi i casi senza influenzare minimamente i modelli (il che è facilmente verificabile; a questo scopo, entrambi i tautomeri di Arg sono mostrati nella Fig.

8a ed entrambi i tautomeri di His sono mostrati nelle Figg. 7b e 8a).

È importante notare che la proiezione preliminare di un segnale ammette l'imposizione di requisiti funzionali come condizioni formali aggiuntive. Il codice terrestre è noto per essere conservativo rispetto ai requisiti polari (Freeland & Hurst, 1998), ma non alle dimensioni molecolari (Haig & Hurst, 1991). Il segnale nel codice non coinvolge il requisito polare in quanto tale, quindi potrebbe essere utilizzato in una condizione formale parallela per ridurre l'effetto di letture errate. Tuttavia, il segnale coinvolge il numero di nucleoni che è correlato al volume molecolare. Ciò interferisce con il tentativo di rendere il codice conservativo anche rispetto alle dimensioni degli amminoacidi.

Possibile interpretazione. Oltre ad avere la funzione di firma intelligente in quanto tale, il segnale del codice genetico potrebbe anche ammettere interpretazioni sensate del suo contenuto. Senza pretendere di essere corretti, proponiamo qui una nostra versione. Oggi si è tentati di pensare che il corpo principale del messaggio possa risiedere nei genomi (Marx, 1979; si veda anche Hoch & Losick, 1997). Sebbene l'idea del SETI genomico (Davies, 2010) possa sembrare ingenua in considerazione delle

mutazioni casuali, le cose non sono così scontate. Per esempio, un locus con un messaggio potrebbe essere esposto a una selezione purificante attraverso l'accoppiamento con geni essenziali, e ci sono persino possibili prove in tal senso (ibid.). In ogni caso, l'ideogramma sembra fornire un riferimento ai genomi. Così, le basi complementari speculari della breve stringa superiore (Fig. 9a) assomigliano a coppie Watson-Crick; le quattro basi centrali TC|GA e l'asse centrale potrebbero quindi rappresentare il simbolo del DNA genomico stesso. Le basi TATAT laterali (Fig. 9b) potrebbero simboleggiare la sequenza consenso presente nei promotori della maggior parte dei geni. Le sequenze codificanti dei geni si trovano tra i codoni Start e Stop. Viceversa, le regioni non tradotte si trovano tra i codoni Stop e Start dei geni vicini. Pertanto, la stringa di triplette nella Fig. 10c potrebbe simboleggiare le regioni intergeniche ed essere interpretata come l'indirizzo del messaggio genomico.

Anche il sistema numerico privilegiato nel codice potrebbe essere interpretato come un'indicazione di una caratteristica simile nei genomi. Si dice spesso che i genomi conservano le informazioni ereditarie in formato digitale quaternario. Esistono 24 possibili numerazioni dei nucleotidi del DNA con le cifre 0, 1, 2, 3. L'ideogramma sembra suggerire quella corretta: T 0, C 1, G 2, A 3. In questo caso la quadrupletta TCGA (Fig. 9a), letta in senso distinto, rappresenta la sequenza naturale preceduta da zero. I codoni palindromici CCC e TCT (Fig. 10b) diventano rispettivamente un simbolo della simmetria digitale quaternaria 1114 e il radix del sistema corrispondente $0104 = 4$. I codoni AGC, o 3214, correlati a livello traslazionale (Fig. 9b) indicano forse le posizioni nella notazione quaternaria del valore di posto, con gli ordini superiori che vengono prima. La somma delle triplette digitali nella stringa TAG + TAA + AAA + ATG + ATG (Fig. 10c) equivale al numero di nucleotidi del codice $30004 = 192$. Inoltre, T come zero si contrappone al numero di nucleotidi del codice 3214. Inoltre, T come zero si contrappone alle altre tre "cifre" del codice "cifre" del codice scomposto (Fig. 6). Infine, ogni coppia di basi complementari nel DNA è pari a 3, quindi la doppia elica appare numericamente come 333...4, e il codone centrale AAA nella Fig. 10c diventa il simbolo del DNA duplex situato tra i geni. Se questa particolare numerazione abbia una relazione con il messaggio genomico, se esiste, è una questione di ulteriori ricerche.

Vale la pena ricordare che tutti i genomi, nonostante le loro enormi dimensioni e diversità, possiedono una caratteristica universale come il codice genetico stesso. Si tratta della cosiddetta seconda regola di Chargaff. In quasi tutti i genomi - da quelli virali a quelli umani - le quantità di nucleotidi complementari, dinucleotidi e oligonucleotidi superiori fino alla lunghezza di ~ 9 sono bilanciate con buona precisione all'interno di un singolo filamento di DNA (Okamura et al., 2007). A differenza della prima regola di Chargaff, che ha trovato rapidamente una base fisico-

chimica, la seconda regola, con il suo ordine totale, non ha ancora una spiegazione ovvia.

Appendice A. Implementazione molecolare del codice genetico

Qui illustriamo il funzionamento molecolare del codice genetico, che spiega perché rimane invariato per miliardi di anni e, allo stesso tempo, può essere facilmente modificato artificialmente, ad esempio per incorporare un segnale. Per semplicità, tralasciamo i dettagli come la U al posto della T nell'RNA, l'energia dell'ATP, l'accoppiamento wobble, ecc. che non influiscono sulla comprensione del punto principale (per i dettagli si veda, ad esempio, Alberts et al., 2008).

Il primo tipo di molecole alla base del codice genetico è costituito dagli RNA di trasferimento (tRNA). I tRNA sono trascritti come prodotti finali dai geni tRNA nei genomi dalla RNA polimerasi (Fig. A1a; per maggiore precisione, il meccanismo è mostrato per l'amminoacido Ser e il suo codone TCC). Con una lunghezza che varia intorno agli 80 nucleotidi, i trascritti di tRNA si ripiegano in una specifica configurazione spaziale dovuta all'appaiamento di basi tra diverse sezioni dello stesso filamento di RNA, analogamente a quanto avviene tra due filamenti di elica di DNA (Fig. A1b). Ai suoi lati opposti, la molecola di tRNA ripiegata presenta un anticodone non appaiato e l'estremità accettore a cui si legherà l'amminoacido. I tRNA con anticodoni diversi che specificano lo stesso amminoacido (ricordiamo che il codice è ridondante) sono identici nella loro configurazione complessiva. I tRNA che specificano amminoacidi diversi differiscono tra loro per gli anticodoni e per altri punti, quindi hanno configurazioni complessive leggermente diverse. Tuttavia, le estremità dell'accettore sono identiche in tutti i tRNA, quindi per il tRNA stesso non fa differenza quale amminoacido sia legato ad esso, indipendentemente dall'anticodone che ha sul lato opposto. Il processo di legame degli amminoacidi ai tRNA viene eseguito da enzimi proteici chiamati aminoacil-tRNA sintetasi (aaRS, Fig A1b, in basso). Normalmente esistono 20 tipi di aaRS, uno per ogni amminoacido, e sono tradotti da geni appropriati nel genoma. Ognuno di questi enzimi riconosce con grande specificità sia l'amminoacido che lo specifica, sia tutti i tRNA che specificano quell'amminoacido; i tRNA sono riconosciuti principalmente dalla loro configurazione complessiva, non esclusivamente dai loro anticodoni (Fig. A1c). Dopo il legame e un ulteriore controllo, l'aaRS rilascia il tRNA carico di amminoacido per consegnarlo al ribosoma (Fig. A1d). A sua volta, il ribosoma non si preoccupa se il tRNA trasporta un amminoacido specificato dal suo anticodone; controlla solo se

l'anticodone del tRNA corrisponde in modo complementare al codone corrente nell'RNA messaggero (mRNA; Fig. A1e). In caso affermativo, l'amminoacido viene trasferito dal tRNA alla catena peptidica in crescita e il tRNA viene rilasciato per essere riciclato. Se il codone e l'anticodone non coincidono, il tRNA con il suo amminoacido viene rimosso dal ribosoma per essere utilizzato in un secondo momento fino a quando non coincide con il codone sull'mRNA (anche con questo overshoot il ribosoma batterico riesce ad aggiungere ~20 amminoacidi al secondo a una catena peptidica). Il meccanismo descritto si traduce in relazioni tra i codoni dell'mRNA e gli amminoacidi (Fig. A1f) che, raccolti insieme in qualsiasi forma conveniente (una possibilità è mostrata nella Fig. 1a), costituiscono il codice genetico. Il punto chiave in termini di modificabilità del codice genetico è che non vi è alcuna interazione chimica diretta tra i codoni dell'mRNA e gli amminoacidi in nessuna fase. Essi interagiscono attraverso molecole di tRNA e aaRS, che possono essere modificate in modo da riassegnare un codone a un altro amminoacido. A titolo di esempio, le figure A1g-k mostrano un modo semplice di modificare il codice in cui due aminoacidi - Ser e Ala - scambiano due dei loro codoni. È noto che nella maggior parte degli organismi gli anticodoni dei tRNA non sono coinvolti nel riconoscimento da parte degli aaRS cognati per questi amminoacidi (Giegé et al., 1998; il fatto si riflette nella Fig. A1c con SARS che non tocca l'anticodone). Pertanto, i tre nucleotidi del gene tRNA^{Ser} corrispondenti all'anticodone potrebbero essere sostituiti (Fig. A1g), in particolare per ottenere l'anticodone GGC corrispondente al codone GCC nell'mRNA, che normalmente codifica Ala (per ottenere l'anticodone per un codone, o viceversa, si deve applicare la regola della complementarità e invertire la tripletta risultante, poiché i filamenti di DNA/RNA complementari hanno direzioni opposte). In seguito, il SARS continuerà a legare Ser al tRNA^{Ser}, anche se ora ha un nuovo anticodone GGC (Fig. A1h). Se si esegue una procedura analoga con i geni tRNA^{Ala} per produrre tRNA^{Ala} con anticodone GGA, il codice genetico verrebbe modificato: Ser e Ala avrebbero scambiato alcuni dei loro codoni (in realtà due codoni, a causa dell'accoppiamento wobble). Tuttavia, la cellula non sopravviverà a tale intervento, poiché tutti i geni codificanti nel genoma rimangono "scritti" con il codice precedente e dopo la traduzione con il nuovo codice producono tutti proteine non funzionali o al massimo semi-funzionali, con Ala occasionalmente sostituito da Ser e viceversa. Per fissare il nuovo codice in una linea cellulare, è necessario modificare in modo appropriato gli mRNA codificanti per lasciare inalterate le sequenze aminoacidiche delle proteine codificate (Fig. A1i). Ciò avverrebbe automaticamente se tutti i geni codificanti venissero riscritti in tutto il genoma in modo da sostituire i codoni TCC con GCC e viceversa (Fig. A1j); tale operazione è possibile quando i genomi vengono addirittura riscritti da zero (Gibson et al., 2010). A questo punto, le sequenze

aminoacidiche delle proteine rimangono inalterate e una cellula prolifera con il nuovo codice genetico (Fig. A1k).

Ora deve essere chiaro perché il codice genetico è altamente protetto da modifiche casuali. Se si verifica una mutazione nel tRNA o nell'aaRS che porta a una riassegnazione dei codoni, tutti i geni del genoma rimangono scritti con il codice precedente e una cellula esce rapidamente di scena senza progenie. Le probabilità che tale mutazione nel tRNA/aaRS sia accompagnata da mutazioni corrispondenti nei geni codificanti in tutto il genoma, con il risultato di proteine inalterate, sono estremamente ridotte, dato che ci sono decine di codoni di questo tipo in migliaia di geni in un genoma. Pertanto, il macchinario del codice genetico subisce una selezione purificatrice eccezionalmente forte che lo mantiene inalterato per miliardi di anni.

Va ricordato che in realtà il processo di modifica intenzionale del codice è più complicato. Ad esempio, i dettagli del riconoscimento del tRNA da parte degli aaRS variano a seconda della specie di tRNA e dell'organismo, e in alcuni casi l'anticodone è coinvolto, parzialmente o interamente, in questo processo. Tuttavia, questo è evitabile, in linea di principio, con metodi appropriati di ingegneria molecolare. Un altro problema è che le modifiche del codice che lasciano inalterate le proteine possono comunque influenzare il livello di espressione genica (Kudla et al., 2009). Pertanto, potrebbe essere necessario adottare ulteriori misure per ripristinare il modello di espressione con il nuovo codice genetico. Si tratta di questioni tecniche superabili; il punto è che non esistono restrizioni principali per modificare artificialmente il codice in qualsiasi modo desiderato. In effetti, metodi elaborati per modificare la configurazione complessiva del tRNA e/o i siti di riconoscimento dell'aaRS potrebbero consentire non solo di scambiare due amminoacidi, ma di introdurne di nuovi.

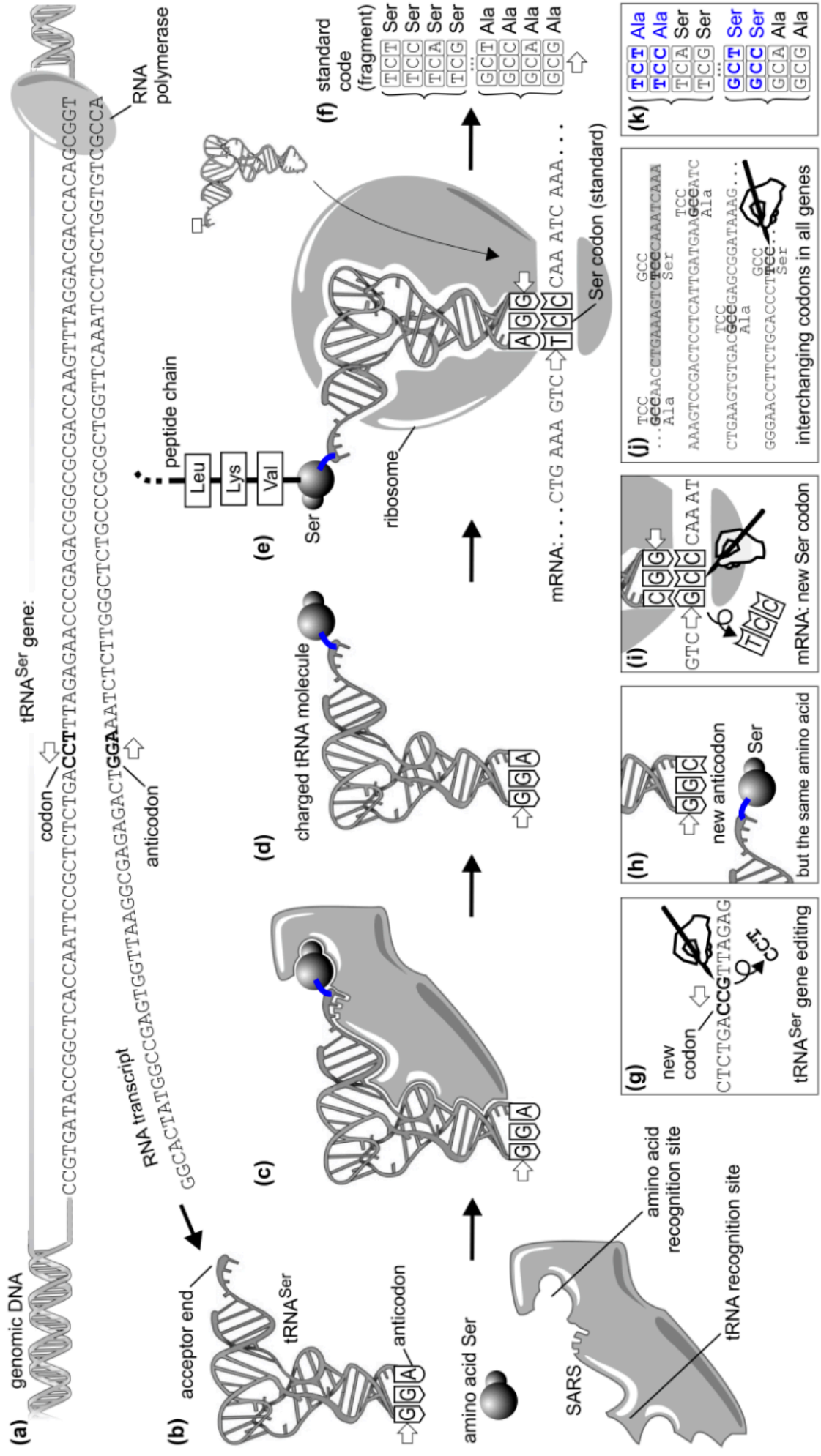


Fig. A1. Meccanismi molecolari del codice genetico (mostrato per il caso dell'amminoacido serina) e un semplice esempio di modifica artificiale. Le frecce di contorno indicano la direzionalità dei filamenti di DNA e RNA, definita dall'orientamento delle loro subunità (designato in biochimica come orientamento $5' \rightarrow 3'$; la replicazione, la trascrizione e la traduzione avvengono solo in quella direzione). (a) Il gene tRNASer (il gene del tRNA che specifica Ser nel codice standard) viene trascritto dalla RNA polimerasi dal DNA genomico. (b) La molecola di tRNASer ripiegata (in alto), la molecola di serina (al centro) e la seril-tRNA sintetasi (SARS, un aaRS cognato per l'amminoacido serina; in basso). (c) SARS riconosce sia la serina che il tRNASer e li lega insieme. (d) SertRNASer rilasciato da SARS e pronto per essere consegnato al ribosoma. (e) Il processo di sintesi del peptide nel ribosoma (come esempio, è mostrato l'mRNA con il frammento genico della SARS stessa). (f) Il frammento di codice genetico risultante (è mostrato anche il gruppo Ala, che sarà utilizzato in un esempio successivo). (g)-(k). Un modo semplice di modificare il codice genetico. La sequenza ombreggiata in (j) corrisponde alla regione mostrata in (e).

Appendice B. Test statistico

È opportuno chiedersi se i pattern presentati siano solo un artefatto della pesca dei dati. Per valutare ciò, si potrebbero confrontare i volumi di informazione dell'insieme di dati in sé (V_0) e dell'insieme di pattern all'interno di tale insieme (V_p). L'artefatto della pesca dei dati potrebbe essere definito come il caso in cui $V_p \ll V_0$. Come illustrato nell'Appendice E, l'insieme di pattern presentato può essere descritto con un sistema di equazioni diofantine, in cui i numeri di nucleoni degli amminoacidi fungono da incognite. Dato l'insieme degli amminoacidi canonici (la gamma di valori possibili per le incognite), questo sistema è completamente definito: ha un'unica soluzione che risulta essere la mappatura effettiva del codice (ciò implica anche che non ci sono più modelli algebricamente indipendenti dello stesso tipo nel codice). Quindi, $V_p = V_0$, quindi l'insieme dei pattern impiega interamente la capacità informativa del codice, dimostrando di rappresentare una caratteristica intrinseca al codice stesso, piuttosto che un artefatto della pesca dei dati.

Ci si potrebbe chiedere quanto sia probabile che un simile insieme di pattern appaia casualmente nel codice genetico. Poiché questa domanda implica che l'attuale mappatura del codice sia stata modellata da processi naturali, è più appropriato chiedersi quanto sia probabile che tale insieme di pattern appaia per caso in determinate condizioni che riflettono presumibili percorsi evolutivi. Abbiamo testato entrambe le versioni dell'ipotesi nulla ("i pattern sono dovuti solo al caso" e "i pattern

sono dovuti al caso accoppiato a presunti percorsi evolutivi"). I risultati sono dello stesso ordine di grandezza; descriviamo solo la versione con percorsi evolutivi presumibili. In questo test sono stati imposti ai codici generati al computer tre percorsi che riflettono le speculazioni predominanti sull'evoluzione del codice:

(1) La ridondanza deve essere in media simile a quella del codice reale. Si ritiene che ciò sia dovuto alle specifiche interazioni tra ribosoma, mRNA e tRNA (Novozhilov et al., 2007). Inoltre, abbiamo tenuto conto della possibile dipendenza della probabilità che una famiglia di codoni rimanga intera o si divida dal tipo delle sue prime due basi. Ciò deriva dalla differenza di termostabilità tra le coppie codone-anticodone arricchite con basi forti (G e C) e quelle arricchite con basi deboli (A e T) (Lagerkvist, 1978). Per questo motivo, la probabilità per una famiglia di quattro codoni con doppietti forti principali di specificare un singolo amminoacido è stata adottata pari a 0,9, per quelli con doppietti deboli - 0,1, e per i doppietti misti è stata di 0,5. Ognuno dei 20 amminoacidi e degli Stop viene reclutato almeno una volta; pertanto i codici con meno di 21 blocchi generati vengono scartati. Successivamente, i blocchi sono stati popolati in modo casuale con amminoacidi e Stop.

(2) Riduzione dell'effetto delle mutazioni/mutazioni dovute alla selezione naturale. La funzione di costo per il requisito polare è stata adottata da Freeland & Hurst (1998), tenendo conto delle distorsioni dovute a trasformazioni e errori di traduzione (si veda anche Novozhilov et al., 2007). Solo i codici che avevano un valore della funzione di costo inferiore a $\varphi_0 + \sigma$, dove φ_0 è il valore per il codice universale e σ è la deviazione standard per tutti i codici casuali filtrati attraverso la condizione precedente, sono stati passati oltre.

(3) Piccolo scostamento dall'equilibrio citoplasmatico (vedi Appendice D). Come sostenuto da Downes & Richardson (2002), questo equilibrio potrebbe riflettere percorsi evolutivi che ottimizzano la distribuzione della massa nelle proteine. Con C che sta per tutti i nucleoni delle catene laterali nel codice e B per tutti i nucleoni dei residui del blocco, il valore $\delta = (C - B)/(C + B)$ si distribuisce approssimativamente in modo normale con $\mu = 0,043$ e $\sigma = 0,024$ (sotto la prima condizione descritta sopra). Sono stati considerati solo i codici che avevano δ nell'intervallo $0 \pm \sigma$, centrato sul valore del codice standard. Poiché questo intervallo corrisponde a codici in cui predominano gli amminoacidi più piccoli ("primi"), questa condizione riflette anche la presumibile storia dell'espansione del codice (Trifonov, 2000; Wong, 2005).

La variabile casuale in questione è il numero di pattern indipendenti dello stesso tipo in un codice. Ovviamente, più pattern di questo tipo vengono osservati in un codice,

meno probabile è tale osservazione. Probabilmente, una buona approssimazione sarebbe una distribuzione binomiale, dato che, ad esempio, un equilibrio di nucleoni può essere considerato come un processo di Bernoulli: in una data disposizione l'equilibrio è "acceso" o "spento", dove la probabilità di "acceso" è molto più piccola di quella di "spento". Tuttavia, le probabilità per le bilance in disposizioni diverse possono essere diverse, soprattutto in base alle condizioni imposte. La situazione è ancora più complessa con le simmetrie degli ideogrammi: la simmetria non è solo "on" o "off", ma è anche caratterizzata dalla lunghezza della stringa e dal numero di tipi di nucleotidi coinvolti. Pertanto, non applichiamo alcuna approssimazione, ma utilizziamo un approccio bruteforce per trovare distribuzioni per punteggi opportunamente definiti per i modelli. La prolina è stata considerata con un nucleone trasferito dalla sua catena laterale al suo blocco (si noti che, poiché la chiave di attivazione è applicata universalmente, il codice reale e il codice con la chiave applicata sono statisticamente equivalenti).

Bilanci dei nucleoni. Gli schemi aritmetici del codice standard sono tutti dello stesso tipo: uguaglianza delle somme dei nucleoni + la loro notazione decimale distintiva + almeno una delle tre trasformazioni (tranne il caso scomposto). La ricerca di un codice casuale con pochi schemi di questo tipo si è rivelata dispendiosa in termini di tempo, quindi i requisiti sono stati notevolmente semplificati. Sono state considerate solo le uguaglianze tra nucleoni, senza che fosse richiesta una notazione distintiva in un sistema numerico. La presenza di trasformazioni è stata richiesta solo nella disposizione di Gamow per i codoni con basi identiche e uniche, poiché le trasformazioni agiscono in primo luogo lì, non come compagni di un'altra logica di ordinamento. Inoltre, per semplicità, sono stati considerati solo i modelli globali; le caratteristiche "locali", come il triplice equilibrio della Fig. 5b, non sono state controllate.

I codici alternativi potrebbero avere equilibri in disposizioni e combinazioni diverse da quelle del codice reale. Contrariamente a quanto potrebbe sembrare, non esistono molti modi di disporre il codice basati su una logica semplice e con un minimo di arbitrarietà. Ad esempio, oltre all'ordinamento di Gamow, nei primi tentativi di dedurre teoricamente il codice sono state proposte diverse altre disposizioni (cfr. Hayes, 1998). Una di queste è nota come "codice senza virgole" (Crick et al., 1957). Tuttavia, a differenza dell'ordinamento di Gamow, questa e altre disposizioni proposte non consentono di "congelare" completamente gli elementi del codice, lasciando un ampio grado di arbitrarietà. In definitiva, nel test sono state prese in considerazione le seguenti disposizioni:

-divisioni basate sulla ridondanza

-divisioni basate sulle posizioni nei codoni (alternando tutte le combinazioni come S o W in prima posizione, R o Y in seconda posizione, ecc;)

-ordinamenti basati sulla composizione nucleotidica dei codoni (alternando tutte le combinazioni di condizioni di "congelamento" e logica di divisione); - ordinamenti basati sulla scomposizione dei codoni in basi (alternando tutte le combinazioni dei quattro insiemi nucleotidici).

Inoltre, i primi due tipi possono essere disposti con codoni a dimensione piena o contratta. È stato verificato anche l'unico equilibrio possibile della rappresentazione peptidica (Appendice D). In totale, sono stati controllati 160 potenziali equilibri (sia del tipo catena-catena che del tipo blocco-catena) in tutte queste disposizioni. Sono state prese precauzioni per ignorare le dipendenze aritmetiche, poiché per alcune versioni del codice alcuni equilibri sono banalmente soddisfatti se se ne verificano pochi altri. È stato adottato un semplice schema di punteggio: il punteggio di un codice è il numero di uguaglianze nucleoniche algebricamente indipendenti che si verifica in tutte le disposizioni. In questo schema la versione semplificata degli schemi aritmetici del codice standard ha un punteggio di 7. Una stima al computer mostra che la probabilità che un codice abbia per caso un punteggio non inferiore a 7 nelle condizioni imposte è $p_1 = 1,5 \times 10^{-8}$ (Fig. B1a).

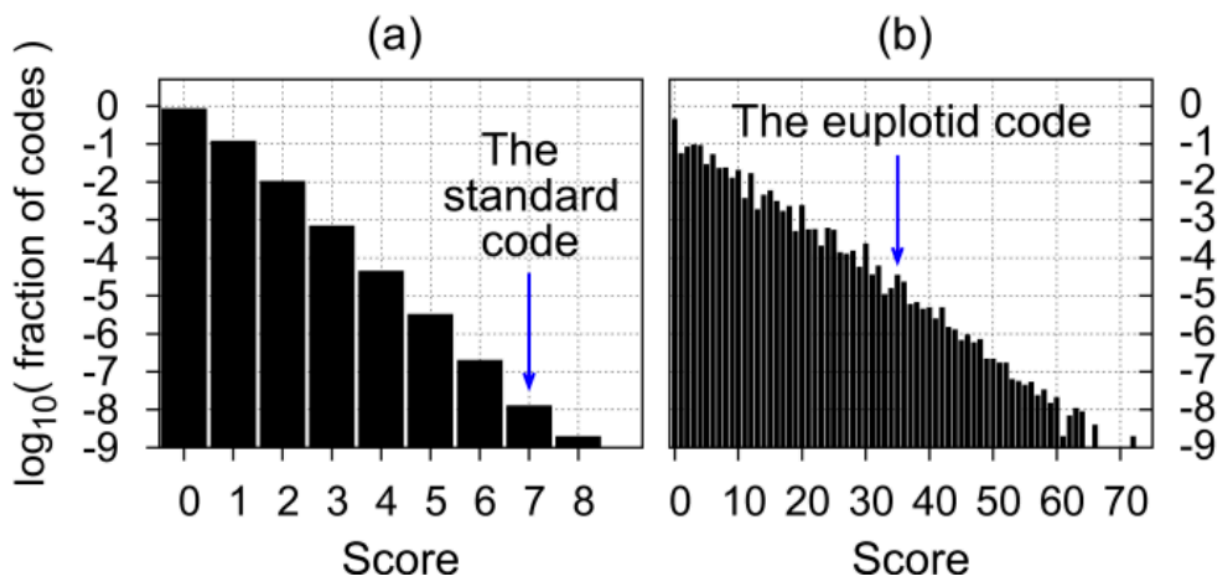


Fig. B1. Distribuzione dei codici varianti in base ai loro punteggi per (a) le uguaglianze dei nucleoni e (b) le simmetrie degli ideogrammi. La dimensione del campione in entrambi i casi è di un miliardo di codici.

Simmetrie dell'ideogramma. Si potrebbe costruire un ideogramma per ogni codice variante nello stesso modo mostrato nella Fig. 7 (tuttavia, non è richiesto che le famiglie intere e divise siano collegate con qualsiasi trasformazione). Esistono altri modi possibili per costruire un ideogramma utilizzando serie di codoni contratte (gli ideogrammi basati su codoni a grandezza naturale soffrono di ambiguità). Ad esempio, i numeri dei nucleoni e della ridondanza potrebbero essere disposti nella stessa direzione, anziché in modo antisimmetrico. Un altro modo è quello di dividere il codice per posizioni nei codoni (ad esempio, R o Y nella prima posizione; tuttavia, questi ideogrammi sono più semplici perché due delle quattro stringhe superiori sono sempre binarie, mentre negli ideogrammi basati sulla ridondanza tutte le stringhe sono, in generale, quaternarie). In totale, sono state costruite 9 versioni di ideogrammi per ogni codice e sono state controllate le simmetrie. Ovvero, ognuna delle quattro stringhe è stata controllata per M, M + I, T, T + I, dove M e T stanno per simmetrie speculari e di traslazione e I denota inversioni di coppia di tutti e tre i tipi. Per ogni simmetria una stringa di lunghezza L riceve il punteggio L/2, se contiene solo due tipi di basi (o se la simmetria vale solo nella rappresentazione binaria RY, SW o KM), e L, se contiene tre o tutti e quattro i tipi di basi. Sono state considerate solo le simmetrie delle stringhe intere (in questo caso non sono state rilevate le simmetrie multiple che organizzano parti diverse di una stringa, come nella Fig. 9b; l'intera stringa nella Fig. 9b, tuttavia, è simmetrica a specchio nella rappresentazione KM). Per ogni posizione ambigua (due serie vicine con numero di nucleoni uguale) è stata introdotta la penalità L/3. Le simmetrie semantiche e gli equilibri degli amminoacidi tradotti non sono stati controllati. Infine, se almeno una delle quattro stringhe non presenta nessuna delle simmetrie, il punteggio viene diviso per 2. Il codice euplotid ha un punteggio di 35 in questo schema: 8 per M + I(TA, CG) e 4 per T_{RY} nella stringa corta superiore, 4 per M_{RY} nella stringa corta centrale, 8 per M_{KM} nella stringa lunga superiore, 16 per M nella stringa lunga centrale, penalità $-16/3 \approx -5$ per Lys e Gln (anche se in questo caso il loro scambio non riguarda né M_{KM} nella stringa superiore, né M in quella centrale). Una stima al computer mostra che la probabilità che un codice abbia per caso un punteggio non inferiore a 35 nelle condizioni imposte è $p_2 = 9,4 \times 10^{-5}$ (Fig. B1b).

Abbiamo anche controllato le trasformazioni nelle bisezioni di Rumer dei codici generati, poiché queste trasformazioni sono servite come principio guida per l'estrazione del segnale nel codice reale. In base alle condizioni imposte, la probabilità per un codice casuale di avere un numero uguale di famiglie intere e di famiglie divise, che sono inoltre collegate con una qualsiasi delle tre possibili trasformazioni, è risultata essere di $4,6 \times 10^{-2}$. Dato che una trasformazione ha luogo, le altre due potrebbero essere distribuite tra i codoni nei rapporti 8:0 ($p = 0,125$), 4:4 ($p = 0,375$) o

2:6 ($p = 0,5$). Per il codice reale questo rapporto è 4:4 (vedi Fig. 2a), quindi alla fine $p_3 = 1,7 \times 10^{-2}$.

Come suggerito da uno studio computazionale separato, l'influenza reciproca dei tre tipi di pattern è trascurabile, quindi la probabilità totale che un segnale (molto semplificato) si verifichi casualmente in un singolo codice nelle condizioni imposte è $p_1 p_2 p_3 = 2,4 \times 10^{-14}$. Poiché il codice simmetrico a ridondanza non deve essere trovato in natura per rivelare l'ideogramma, il valore P finale non si discosterà molto da questo valore.

Questo risultato fornisce probabilità per un tipo specifico di modelli - uguaglianze di nucleoni, simmetrie di ideogrammi e trasformazioni. Tuttavia, la verifica dell'ipotesi di un segnale intelligente dovrebbe prendere in considerazione anche modelli di altro tipo, purché soddisfino i requisiti indicati nell'introduzione. Dopo aver analizzato la letteratura sul codice genetico, riteniamo che i numeri di nucleoni e di ridondanza siano i migliori candidati per i "numeri ostensivi". Accettiamo tuttavia che ci possano essere altre possibilità e che il valore P ottenuto debba essere considerato come un'approssimazione approssimativa (si tenga conto anche delle semplificazioni del test). Tuttavia, ammettiamo che non ci sono abbastanza candidati per i "numeri ostensivi" e i corrispondenti insiemi di modelli (algebricamente definiti) per compensare il piccolo valore P ottenuto e portarlo vicino al livello di significatività.

Appendice C. Simmetrie digitali dei sistemi numerali posizionali

La simmetria digitale descritta nel testo principale per il sistema decimale è legata a un criterio di divisibilità che potrebbe essere utilizzato per eseguire in modo efficace le checksum. Consideriamo il numero 27014319417 come esempio. La cornice di lettura della tripletta divide questo numero in triplette digitali 270, 143, 194, 170 (si può scegliere una qualsiasi delle tre cornici di lettura; gli zeri vengono aggiunti ai lati per formare triplette complete). La somma di queste triplette è pari a 777. La sua notazione distintiva indica che il numero originario è divisibile per 037. Nei numeri a quattro cifre che compaiono durante le sommatorie, le cifre delle migliaia vengono trasferite alle cifre dell'unità. Se la notazione della somma risultante non è distintiva, aggiungere o sottrarre 037 una volta. La successiva notazione distintiva confermerà la divisibilità del numero originale per 037, mentre la sua assenza la smentirà. Così, le altre due cornici per il numero esemplare producono:

$$002 + 701 + 431 + 941 + 700 = 2775 \rightarrow 002 + 775 = 777; 027 + 014 + 319 + 417 = 777.$$

Questo criterio si applica a numeri di qualsiasi lunghezza e richiede un registro con solo tre posizioni. Procedendo lungo una notazione lineare, tale registro somma le terzine digitali e trasferisce le cifre delle migliaia alle cifre dell'unità.

La stessa simmetria digitale delle triplette e il relativo criterio di divisibilità esistono in tutti i sistemi numerici con radix q che soddisfano il requisito $(q - 1)/3 = \text{intero}$. Il numero primo correlato alla simmetria in questi sistemi si trova come $111_q/3$. Così, la caratteristica esiste nel sistema quaternario ($q = 4$) con il numero primo 7 (013_4), nel sistema settenario ($q = 7$) con il numero primo 19 (0257), nel sistema decimale ($q = 10$) con il numero primo 037, nel sistema con $q = 13$ e il numero primo 61 (049_{13}), e così via. La simmetria digitale del sistema quaternario è illustrata nella Fig. C1.

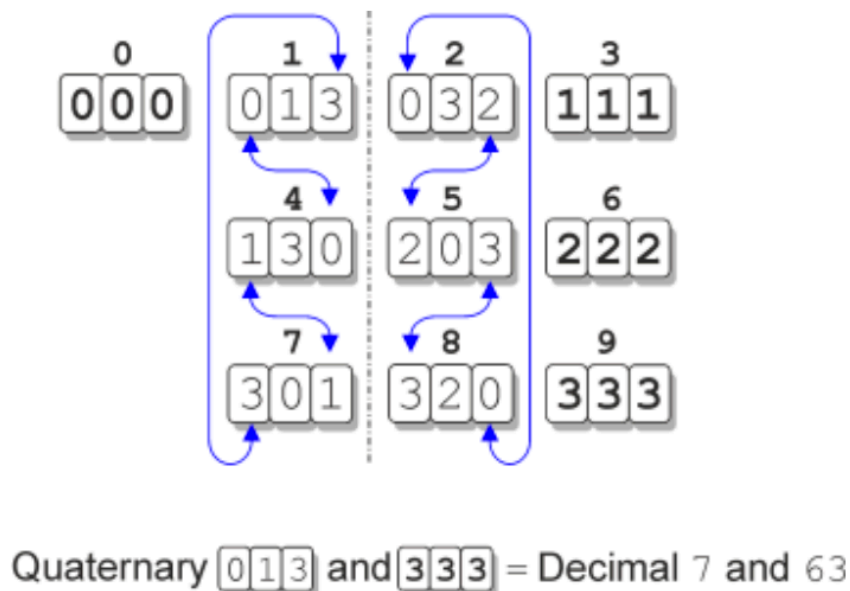


Fig. C1. Analogamente al sistema decimale, anche il sistema quaternario presenta una simmetria delle triplette digitali, dove 7 (013_4) agisce al posto di 037.

Appendice D. L'equilibrio citoplasmatico

La figura D1 rappresenta l'intero codice genetico come un peptide. Ogni amminoacido è inserito in questo peptide tante volte quante ne compaiono nel codice standard. I residui di blocco degli amminoacidi costituiscono la spina dorsale del peptide. Il polimero risultante è lungo 61 amminoacidi. Se i suoi terminali N e C vengono eliminati chiudendo il peptide in un anello, la sua spina dorsale e le sue catene laterali appaiono esattamente bilanciate. Questa caratteristica è comune alle proteine naturali: la loro massa è distribuita in modo approssimativamente uguale tra spina

dorsale e catene laterali (Downes & Richardson, 2002). Ciò implica automaticamente che la frequenza degli amminoacidi nelle proteine naturali è correlata alla loro abbondanza nel codice genetico (vedi dati in Gilis et al., 2001).

In questo equilibrio non viene scartata solo la chiave di attivazione, ma le molecole di amminoacidi vengono considerate come appaiono nell'ambiente citoplasmatico (dove le catene laterali di alcune di esse sono ionizzate). Per questi motivi l'equilibrio mostrato in Fig. D1 viene definito naturale o citoplasmatico. Tuttavia, la forma insolita del peptide (anche se i peptidi circolari sono rari in natura, vedi Conlan et al., 2010) e la distinzione tra blocchi e catene di amminoacidi suggeriscono che l'equilibrio citoplasmatico e gli equilibri "virtuali" mostrati nel testo principale sono probabilmente fenomeni correlati. Probabilmente, questo equilibrio ha lo scopo di convalidare la natura artificiale della chiave di attivazione, dimostrando che solo la prolina reale può mantenere i modelli nell'ambiente naturale. Questo equilibrio è stato trovato da Downes & Richardson (2002) sotto il profilo biologico. Contemporaneamente, Kashkarov et al. (2002) lo hanno trovato con un approccio aritmetico formale.

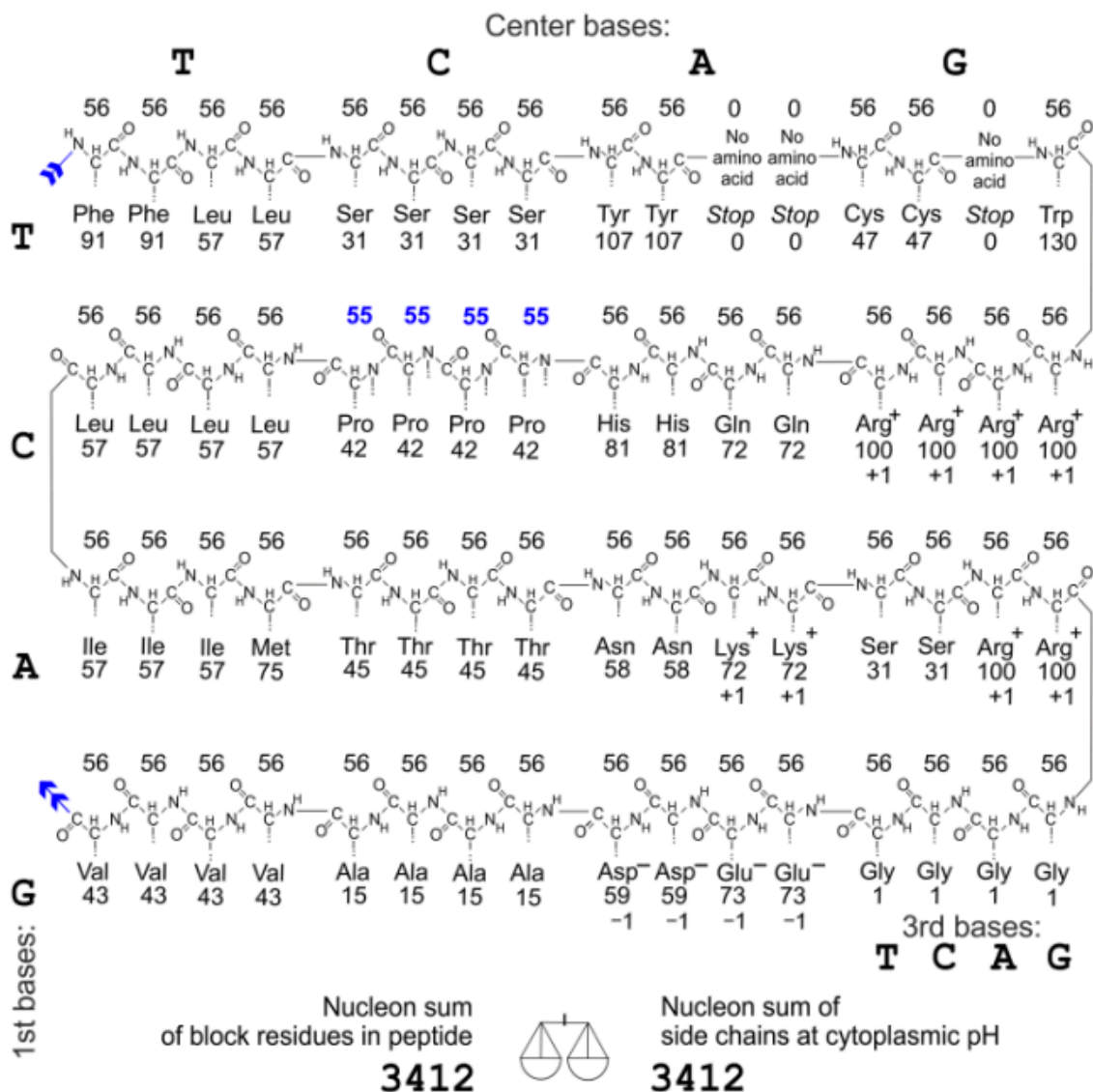


Fig. D1. Amminoacidi del codice genetico standard sotto forma di peptide circolare (l'ordine di sequenza non ha importanza). Il peptide si forma aggregando blocchi standard di amminoacidi in una spina dorsale polimerica. La formazione di ogni legame peptidico libera una molecola d'acqua, riducendo ogni blocco di amminoacidi a 56 nucleoni (55 nella prolina). Asp e Glu perdono un protone ciascuno dalle loro catene laterali a pH citoplasmatico, mentre Arg e Lys guadagnano un protone ciascuno (indicati rispettivamente con -1 e +1). Gli altri amminoacidi sono prevalentemente neutri nell'ambiente citoplasmatico (Alberts et al., 2008). Di conseguenza, la somma dei nucleoni della spina dorsale peptidica è esattamente uguale a quella di tutte le sue catene laterali.

Appendice E. Rappresentazione algebrica del segnale

Descriviamo qui un possibile modo in cui si sarebbe potuta ottenere la mappatura segnale-arborazione. Come dati iniziali, si ha un insieme di 64 codoni e un altro insieme di 20 amminoacidi canonici più Stop. Supponiamo che la mappatura tra questi due insiemi sia sconosciuta e che debba essere dedotta dall'insieme di pattern del segnale. Esistono $\sim 10^{83}$ possibili mappature tra i due insiemi, a condizione che ogni elemento del secondo insieme sia rappresentato almeno una volta. Conoscere l'ideogramma (senza conoscere i numeri dei nucleoni mappati ai singoli codoni) equivale a conoscere la struttura a blocchi del codice. Da ciò deriva la prima parte delle equazioni $ggt = ggc = gga = ggg = ggn$, $ttt = ttc = tty$, ecc. dove i codoni sono usati per indicare le variabili - i numeri di nucleoni sconosciuti delle catene laterali degli amminoacidi. In questo modo, il numero di elementi del primo insieme si riduce essenzialmente da 64 a 24. Ma ci sono ancora ~ 1030 elementi possibili. Ma rimangono ancora ~ 1030 possibili mappature. Ora si potrebbero scrivere le somme dei nucleoni delle Figg. 5-8 e 10 (tralasciando le parti dipendenti dall'algebra e le somme a blocchi standard, dato che abbiamo a disposizione l'insieme degli amminoacidi canonici; in caso di proiezione dei modelli Stop potrebbe essere assegnato preliminarmente a certi codoni per facilitare le cose con le somme a blocchi):

$$\begin{aligned}
&ggn + gcn + tcn + ccn + gtn + acn + ctn + cgn = 333 \text{ (Fig. 7b);} \\
&tgy + tga + ath + tar + agy + ttr + aay + gay + car + aar + gar \\
&\quad + cay + tty + agr + tay + atg + tgg = 111 + 999 \text{ (Fig. 7b);} \\
&tty + ttr + tcn + tay + tar + tgy + tga + tgg + ctn + ccn + cay \\
&\quad + car + cgn = 814 \text{ (Fig. 8a);} \\
&tty + ttr + tcn + tay + tar + tgy + tga + tgg + gtn + gcn + gay \\
&\quad + gar + ggn = 654 \text{ (Fig. 8b);} \\
&tty + ttr + ctn + ath + atg + gtn + tgy + tga + tgg + cgn + agy \\
&\quad + agr + ggn = 789 \text{ (Fig. 8b);} \\
&tty + aar + ath + tcn + cay + 2gcn + ctn + tgy + tga + gay + atg \\
&\quad + car + agy = 703 \text{ (Fig. 5a);} \\
&ggn + ccn + ctn + 2acn + tay + tcn + 2gtm + 2cgn + agy + tar \\
&\quad + gay = 703 \text{ (Fig. 5a);} \\
&tty + 2ttr + 3ccn + 2ctn + ath + gtn + 2tcn + acn + gcn + tay \\
&\quad + tgy + cay + cgn = 999 \text{ (Fig. 5b);}
\end{aligned}$$

$$\begin{aligned}
&2aay + aar + tar + car + gar = 333 \text{ (Fig. 5b);} \\
&3ggn + tgg + cgn + agr = 333 \text{ (Fig. 5b);} \\
&ath + acn + agr + gtn + gcn + gar = 333 \text{ (Fig. 5b);} \\
&tty + 2ctn + 2tcn + ccn + 2aay + tar + ath + car + acn + 2ggn \\
&\quad + tgg + gtn + cgn + gcn = 888 \text{ (Fig. 5c);} \\
&5tty + 4ttr + 5ctn + 4ath + atg + 5gtm + 5tcn + ccn + acn + gcn \\
&\quad + 3tay + 2tar + cay + aay + gay + 3tgy + tga + tgg + cgn \\
&\quad + agy + ggn = 666 + 999 \times 2 \text{ (Fig. 6b);} \\
&2tar + aar + 2atg = 222 \text{ (Fig. 10d);} \\
&agy + 2aar + tgh = 222 \text{ (Fig. 10e).}
\end{aligned}$$

L'equilibrio citoplasmatico non viene preso in considerazione in quanto non ha alcuna connessione algebrica con questo sistema a causa della chiave di attivazione. Vi sono inoltre ulteriori disuguaglianze fornite dall'ideogramma (Fig. 7a):

$$\begin{aligned}
&ggn \leq gcn \leq tcn \leq ccn \leq gtn \leq acn \leq ctn \leq cgn; \\
&tgh \leq ath; \\
&tar \leq agy \leq ttr \leq aay \leq gay \leq car \leq aar \leq gar \leq cay \leq tty \leq \\
&agr \leq tay; \\
&atg \leq tgg.
\end{aligned}$$

Infine, $tgh = tgy$ per tenere conto di due versioni del codice. In totale, ci sono 26 incognite, 16 equazioni e 20 disuguaglianze. In genere, questi sistemi di equazioni e disequazioni diofantine hanno più soluzioni. Poiché qui ci interessa ottenere la mappatura del codice dati i modelli e l'insieme fisso di amminoacidi canonici più Stop, la soluzione va cercata sul dominio frammentario di numeri interi e zero $\{0, 1, 15, 31, 41, 43, 45, 47, 57, 58, 59, 72, 73, 75, 81, 91, 100, 107, 130\}$. In questo caso, l'analisi del sistema con un qualsiasi sistema di algebra informatica in grado di trattare espressioni diofantine mostra che questo sistema ha un'unica soluzione che coincide con l'effettiva mappatura dei numeri nucleonici sui codoni: $tty = 91$, $ggn = 1$, $tga = 0$, $ath = 57$, ecc. Rimangono però diverse mappature per gli amminoacidi, poiché due delle radici - 57 e 72 - rappresentano due amminoacidi ciascuna. Questa ambiguità viene eliminata quando si prendono in considerazione anche i modelli all'interno delle catene laterali (Figg. 7b e 8a). In seguito, la mappatura effettiva del codice viene dedotta senza ambiguità dal sistema algebrico dei pattern. In effetti, l'analisi mostra che si ottiene una soluzione univoca anche se la restrizione del dominio frammentario viene applicata solo ad alcune delle incognite. In un altro approccio (shCherbak, 2003) la soluzione univoca si ottiene solo con poche assunzioni sull'insieme degli amminoacidi.

Ringraziamenti

Lo studio è stato parzialmente finanziato dal Ministero dell'Istruzione e della Scienza della Repubblica del Kazakistan. La ricerca è stata promossa dal professor Bakytzhan T. Zhumagulov dell'Accademia nazionale di ingegneria della Repubblica del Kazakistan. Parte della ricerca è stata svolta durante V.I.S. ha soggiornato presso il Max-Planck-Institut für biophysikalische Chemie (Göttingen, Germania) su gentile invito del professor Manfred Eigen. V.I.S. esprime un ringraziamento speciale a Ruthild Winkler-Oswatitsch per il suo prezioso aiuto e le sue cure. M.A.M. riconosce il supporto dell'amministrazione dell'Istituto Astrofisico Fesenkov. Gli autori sono grati al professor Paul C.W. Davies, Felix P. Filatov, Vladimir V. Kashkarov, Artem S. Novozhilov, Denis V. Tulinov, Artem N. Yermilov e Denis V. Yurin per le critiche obiettive e le proficue discussioni sul manoscritto. Apprezziamo profondamente la Redazione di Icarus per l'organizzazione della peer review e i tre revisori per i loro commenti che hanno portato al miglioramento del manoscritto.

Contributi degli autori

V.I.S. ha ideato ed eseguito la ricerca, sviluppando le arti grafiche. V.I.S. e M.A.M. hanno analizzato i dati, introdotto l'interpretazione della chiave di attivazione, delineato la struttura dell'articolo. M.A.M. ha eseguito test statistici e analisi algebriche, ha scritto il manoscritto.

SEI ALTROVE EDIZIONI