Supplementary Tables

Supplementary Table 1. This table is analogous to Table 1 but the data on all 20 amino acids are shown. For each amino acid, beneath the D value, the analogous value obtained with a correction taking into account possible substitutions on the path from the outgroup to the common ancestor of the sister genomes (see Supplementary Method), is shown. Asterisks mark D values which deviate significantly from 0 ($P < 0.01$; no correction for multiple tests was applied).

Supplementary Table 1 **Changes of frequencies of all amino acids in the 15 taxa**

Taxon	Substitutions to and from an amino acid				
	Cys	Met	His	Ser	Phe
Hominidae	137 / 72 +0.31* +0.30*	324 / 204 +0.23* +0.14*	297 / 206 +0.18* +0.10	633 / 586 +0.04 +0.02	231 / 115 +0.34* +0.32*
Muridae	2547 / 1729 +0.19* +0.18*	5920 / 4205 +0.17* +0.06*	5702 / 4752 +0.09* -0.00	19904/18108 +0.05* -0.01	4461 / 3895 +0.07* +0.09*
Saccharomyces	778 / 430 +0.29* +0.26*	1768 / 1455 +0.10* +0.07*	1715 / 1606 +0.03 -0.00	7776 / 7567 +0.01 +0.01	1824 / 1546 +0.08* +0.10*
Pyrococcus	35 / 21 +0.25 +0.34*	772 / 426 +0.29* +0.16*	277 / 244 +0.06 +0.08	1756 / 1222 +0.18* +0.05*	680 / 598 +0.06 +0.07*
Escherichia	143 / 57 +0.43* +0.40*	238 / 155 +0.21* +0.20*	284 / 194 +0.19* +0.17*	738 / 549 +0.15* +0.13*	151 / 122 +0.11 +0.12
Salmonella	67 / 11 +0.79* +0.59*	127 / 81 +0.22* +0.16*	122 / 74 +0.25* +0.10	365 / 216 +0.26* +0.13*	84 / 57 +0.19 +0.19
Buchnera	75 / 35 +0.36* +0.28*	166 / 102 +0.25* +0.12	191 / 105 +0.29* +0.18*	654 / 587 +0.05 -0.02	249 / 164 +0.21* +0.13*
Vibrio	4 / 0 +1.00 +0.46	22 / 15 +0.19 +0.11	16 / 10 +0.23 +0.08	64 / 52 +0.10 -0.03	18 / 17 +0.03 +0.08
Pseudomonas	501 / 339 +0.19* +0.16*	2985 / 1090 +0.47* +0.25*	1737 / 1252 +0.16* +0.11*	6289 / 4277 +0.19* +0.08*	1956 / 1623 +0.09* +0.09*
Bordetella	89 / 14 +0.73* +0.71*	104 / 66 +0.22* +0.22*	113 / 75 +0.20* +0.20*	301 / 153 +0.33* +0.32*	50 / 51 -0.01 -0.01
Helicobacter	15 / 11 +0.15 +0.10	48 / 20 +0.41* +0.15	55 / 24 +0.39* +0.12	93 / 75 +0.11 -0.05	31 / 28 +0.05 +0.05
Chlamydia	107 / 59 +0.29* +0.10	212 / 137 +0.22* +0.11*	249 / 150 +0.25* +0.08	911 / 761 +0.09* -0.02	238 / 160 +0.20* +0.14*
Bacillus	235 / 126 +0.30* +0.25*	960 / 722 +0.14* +0.10*	910 / 739 +0.10* +0.05	2044 / 1861 +0.05* +0.03	802 / 692 +0.07* +0.08*
Streptococcus	20 / 3 +0.74* +0.44	27 / 26 +0.02 +0.03	38 / 28 +0.15 -0.01	114 / 61 +0.30* +0.15	28 / 21 +0.14 +0.15
Staphylococcus	10 / 1 +0.83 +0.60	32 / 23 +0.16 +0.12	30 / 25 +0.09 -0.01	97 / 75 +0.13 +0.05	44 / 31 +0.17 +0.18
Average D S.E. of D	+0.452 0.07	+0.219 0.03	+0.178 0.03	+0.135 0.03	+0.120 0.02

Supplementary Table 1 (continued)

Taxon	Substitutions to and from an amino acid				
	Asn	Thr	Ile	Val	Arg
Hominidae	340 / 277 +0.10 +0.04	504 / 524 -0.02 -0.04	525 / 450 +0.08 +0.03	725 / 674 +0.04 +0.04	570 / 670 -0.08* -0.08*
Muridae	8238 / 8956 -0.04* -0.03*	15045/12720 +0.08* -0.00	10458/11678 -0.06* -0.03*	16810/14108 +0.09* +0.03*	13228/ 9275 +0.18* +0.12*
Saccharomyces	5421 / 6069 -0.06* -0.08*	5246 / 5849 -0.05* -0.06*	6044 / 6019 +0.00 +0.01	6475 / 5638 +0.07* +0.02*	4755 / 3023 +0.22* +0.17*
Pyrococcus	1107 / 889 +0.11* +0.05*	1081 / 970 +0.05 +0.05*	3620 / 3555 +0.01 -0.00	3357 / 3185 +0.03 -0.00	2994 / 1932 +0.22* +0.14*
Escherichia	491 / 262 +0.30* +0.28*	638 / 564 +0.06 +0.05	638 / 364 +0.27* +0.27*	686 / 643 +0.03 +0.02	298 / 374 -0.11* -0.10*
Salmonella	198 / 140 +0.17* +0.11	295 / 218 +0.15* +0.05	336 / 211 +0.23* +0.16*	362 / 343 +0.03 +0.02	155 / 171 -0.05 +0.00
Buchnera	452 / 587 -0.13* -0.11*	350 / 275 +0.12* +0.01*	839 / 1140 -0.15* -0.08*	831 / 558 +0.20* +0.07*	243 / 149 +0.24* +0.17*
Vibrio	34 / 19 +0.28 +0.12	56 / 34 +0.24 +0.09	74 / 64 +0.07 +0.01	82 / 74 +0.05 +0.04	25 / 20 +0.11 +0.13
Pseudomonas	2912 / 1635 +0.28* +0.14*	4589 / 2870 +0.23* +0.10*	5705 / 3838 +0.20* +0.08*	6328 / 6057 +0.02 +0.02	2389 / 4267 -0.28* -0.09*
Bordetella	98 / 61 +0.23* +0.22*	287 / 183 +0.22* +0.22*	140 / 74 +0.31* +0.31*	301 / 260 +0.07 +0.07	130 / 237 -0.29* -0.29*
Helicobacter	74 / 67 +0.05 -0.05	64 / 65 -0.01 -0.03	149 / 225 -0.20* -0.06	211 / 164 +0.13 -0.01	65 / 55 +0.08 +0.19*
Chlamydia	323 / 306 +0.03 +0.01	497 / 393 +0.12* +0.04	1003 / 1041 -0.02 +0.02	1105 / 906 +0.10* -0.00	470 / 303 +0.22* +0.13*
Bacillus	1843 / 1856 -0.00 -0.02	1798 / 1819 -0.01 -0.02	3095 / 2634 +0.08* +0.07*	2812 / 2804 +0.00 -0.02	1211 / 848 +0.18* +0.16*
Streptococcus	69 / 52 +0.14 +0.13	77 / 71 +0.04 +0.00	130 / 90 +0.18* +0.15	123 / 121 +0.01 -0.02	46 / 44 +0.02 +0.09
Staphylococcus	91 / 82 +0.05 -0.01	82 / 76 +0.04 +0.02	153 / 129 +0.09 +0.09	128 / 106 +0.09 +0.03	33 / 25 +0.14 +0.17
Average D S.E. of D	+0.101 0.04	+0.085 0.02	+0.072 0.04	+0.063 0.01	+0.052 0.05

Supplementary Table 1 (continued)

Taxon	Substitutions to and from an amino acid				
	Gln	Trp	Leu	Tyr	Asp
Hominidae	359 / 291 +0.11* +0.09	61 / 25 +0.42* +0.44*	394 / 397 -0.00 +0.05	96 / 114 -0.09 -0.05	197 / 308 -0.22* -0.18*
Muridae	6910 / 8265 -0.09* -0.09*	732 / 589 +0.11* +0.20*	10479/11447 -0.04* +0.05*	2538 / 2750 -0.04* +0.05*	8799 / 8107 +0.04* +0.02*
Saccharomyces	1803 / 2568 -0.18* -0.14*	62 / 99 -0.23* -0.04	3375 / 3706 -0.05* +0.03*	1097 / 1081 +0.01 +0.07*	4141 / 4875 -0.08* -0.08*
Pyrococcus	450 / 294 +0.21* +0.11*	38 / 59 -0.22 -0.06	1810 / 2141 -0.08* -0.03	383 / 502 -0.13* -0.05	1650 / 1170 +0.17* +0.13*
Escherichia	311 / 367 -0.08 -0.09	19 / 34 -0.28 -0.26	436 / 394 +0.05 +0.07	126 / 80 +0.22 +0.25*	391 / 493 -0.12* -0.12*
Salmonella	108 / 146 -0.15 -0.16*	14 / 16 -0.07 +0.05	211 / 230 -0.04 +0.03	77 / 48 +0.23 +0.25*	173 / 264 -0.21* -0.20*
Buchnera	285 / 161 +0.28* +0.15*	5 / 13 -0.44 -0.27	392 / 474 -0.10* +0.01	109 / 137 -0.11 +0.07	287 / 224 +0.12* +0.11*
Vibrio	28 / 29 -0.02 -0.18	1 / 1 +0.00 +0.33	50 / 45 +0.05 +0.14	9 / 15 -0.25 -0.20	54 / 54 +0.00 +0.02
Pseudomonas	4083 / 3428 +0.09* -0.04*	216 / 314 -0.19* -0.02	4237 / 7591 -0.28* -0.13*	1212 / 1302 -0.04 -0.01	3835 / 3696 +0.02 +0.02
Bordetella	102 / 76 +0.15 +0.15	8 / 13 -0.24 -0.24	194 / 113 +0.26* +0.27*	55 / 32 +0.26 +0.26	160 / 189 -0.08 -0.08
Helicobacter	39 / 30 +0.13 -0.06	6 / 1 +0.71 +0.60	73 / 54 +0.15 +0.11	11 / 32 -0.49* -0.22	48 / 48 +0.00 +0.07
Chlamydia	296 / 246 +0.09 -0.10*	13 / 13 +0.00 +0.20	398 / 569 -0.18* -0.04	111 / 144 -0.13 +0.06	361 / 372 -0.02 +0.03
Bacillus	1025 / 1153 -0.06* -0.07*	40 / 53 -0.14 -0.02	1509 / 1576 -0.02 +0.02	485 / 503 -0.02 +0.04	1484 / 1621 -0.04 -0.05*
Streptococcus	38 / 25 +0.21 +0.00	1 / 2 -0.33 -0.33	51 / 44 +0.07 +0.10	13 / 16 -0.10 +0.11	72 / 91 -0.12 -0.12
Staphylococcus	32 / 41 -0.12 -0.19	2 / 0 +1.00 +1.00	68 / 68 +0.00 +0.06	38 / 35 +0.04 +0.11	84 / 103 -0.10 -0.08
Average D S.E. of D	+0.037 0.04	+0.007 0.11	-0.014 0.03	-0.042 0.05	-0.042 0.03

Supplementary Table 1 (continued)

Taxon	Substitutions to and from an amino acid				
	Lys	Gly	Glu	Ala	Pro
Hominidae	336 / 292 +0.07 +0.08	294 / 342 -0.08 -0.02	232 / 386 -0.25* -0.19*	517 / 606 -0.08* -0.07	204 / 437 -0.36* -0.30*
Muridae	6829 /10089 -0.19* -0.11*	8238/8677 -0.03* +0.05*	8056 /11269 -0.17* -0.19*	1559 / 1733 -0.05* -0.03*	7350 / 9883 -0.15* -0.07*
Saccharomyces	4398 / 5742 -0.13* -0.10*	3370/2511 +0.15* +0.17*	4006 / 4554 -0.06* -0.19*	5336 / 4950 +0.04* -0.00	2188 / 2290 -0.02 +0.01
Pyrococcus	2731 / 3768 -0.16* -0.14*	727 / 757 -0.02 +0.13*	1663 / 2836 -0.26* -0.17*	1241 / 1598 -0.13* -0.05*	211 / 416 -0.33* -0.19*
Escherichia	286 / 257 +0.05 +0.06	239 / 334 -0.17* -0.15*	360 / 476 -0.14* -0.13*	531 / 1129 -0.36* -0.35*	138 / 294 -0.36* -0.34*
Salmonella	103 / 118 -0.07 -0.01	126 / 205 -0.24* -0.08	176 / 202 -0.07 -0.04	285 / 519 -0.29* -0.21*	53 / 167 -0.52* -0.36*
Buchnera	437 / 816 -0.30* -0.22*	99 / 149 -0.20* +0.07	327 / 263 +0.11* +0.03	403 / 396 +0.01 -0.02	65 / 124 -0.31* -0.06
Vibrio	37 / 40 -0.04 +0.03	23 / 24 -0.02 +0.19	45 / 70 -0.22 -0.17	45 / 93 -0.35* -0.19*	10 / 21 -0.36 -0.10
Pseudomonas	2994 / 2423 +0.11* -0.00	2759/3186 -0.07* +0.08*	3780 / 5448 -0.18* -0.12*	6623 / 9261 -0.17* -0.11*	1054 / 2287 -0.37* -0.17*
Bordetella	67 / 57 +0.08 +0.09	157 / 247 -0.22* -0.22*	102 / 125 -0.10 -0.09	248 / 575 -0.40* -0.40*	83 / 188 -0.39* -0.38*
Helicobacter	71 / 95 -0.15 -0.14	45 / 42 +0.03 +0.29*	44 / 67 -0.21 -0.15	93 / 119 -0.12 -0.00	14 / 27 -0.32 -0.05
Chlamydia	305 / 609 -0.33* -0.14*	174 / 198 -0.07 +0.15*	307 / 512 -0.25* -0.16*	697 / 790 -0.06 -0.02	116 / 224 -0.32* -0.05
Bacillus	1874 / 2047 -0.04* -0.03	898 / 932 -0.02 +0.06*	1702 / 2258 -0.14* -0.13*	1911 / 2117 -0.05* -0.05*	342 / 619 -0.29* -0.22*
Streptococcus	65 / 64 +0.01 +0.02	28 / 50 -0.28 -0.01	61 / 101 -0.25* -0.17	71 / 128 -0.29* -0.19*	13 / 47 -0.57* -0.35*
Staphylococcus	86 / 90 -0.02 -0.01	33 / 53 -0.23 -0.03	81 / 93 -0.07 -0.07	76 / 103 -0.15 -0.11	6 / 47 -0.77* -0.39*
Average D S.E. of D	-0.075 0.04	-0.098 0.03	-0.150 0.03	-0.163 0.04	-0.362 0.04

Supplementary Table 2. This table presents the pattern of long-term gain or loss of amino acids obtained by comparing extant proteins with their reconstructed remote ancestors. Two methods for deep ancestral reconstructions: PAML[1] and EMAPI[2], were applied to two sets of proteins. The first set, employed for reconstructing LUCA proteins, included 32 ribosomal proteins from bacteria (*Bacillus halodurans* C-125, *Deinococcus radiodurans* R1, *Haemophilus influenzae* Rd KW20, *Helicobacter pylori* 26695, *Treponema pallidum* subsp. *pallidum* str. Nichols), archaea (*Archaeoglobus fulgidus* DSM 4304, *Aeropyrum pernix* K1, *Methanocaldococcus jannaschii* DSM 2661, *Sulfolobus solfataricus* P2, *Thermoplasma volcanium* GSS1), and eukaryotes (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*). The construction of the alignments of ribosomal proteins and removal of poorly aligned regions were described previously[3]. Altogether, 2836 sites were analyzed. The second set, which was used for the reconstruction of the ancestral eukaryotic sequences, consisted of 684 conserved proteins from 8 eukaryotic species (*Homo sapiens*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombae*, and *Plasmodium falciparum*). The construction and filtering of these alignments was described previously[4]; 162719 sites were used for the reconstruction.

Obviously, the results of these reconstructions are not very reliable, since PAML and EMAPI produce substantially different values when applied to the same data set. This is hardly surprising because both methods use a time-reversible JTT substitution matrix[5], which is based on the incorrect assumption of detailed equilibrium and uses present-day amino acid composition of proteins to compute substitution probabilities (PAML also assumes constant amino acid composition of proteins).

Supplementary Table 2 **Long-term rates of amino acid gain and loss**

Amino Acid	Average long-term rate of gain/loss†	Ribosomal proteins with ancestors reconstructed by PAML‡	Ribosomal proteins with ancestors reconstructed by EMAPI‡	Eukaryotic proteins with ancestors reconstructed by PAML‡	Eukaryotic proteins with ancestors reconstructed by EMAPI‡
Cys	+0.0055	+0.0034	+0.0066	+0.0058	+0.0061
Met	+0.0063	+0.0059	+0.0055	+0.0062	+0.0077
His	+0.0010	+0.0004	-0.0008	+0.0028	+0.0018
Ser	+0.0116	+0.0138	+0.0297	+0.0020	+0.0009
Phe	+0.0052	+0.0048	+0.0103	+0.0033	+0.0026
Asn	-0.0052	+0.0037	-0.0002	-0.0041	-0.0201
Thr	+0.0069	+0.0031	+0.0116	+0.0043	+0.0088
Ile	-0.0092	+0.0026	-0.0073	+0.0010	-0.0329
Val	-0.0070	-0.0043	-0.0392	+0.0026	+0.0128
Arg	-0.0059	-0.0041	-0.0286	+0.0018	+0.0074
Gln	+0.0103	+0.0057	+0.0193	+0.0037	+0.0123
Trp	+0.0022	+0.0014	+0.0047	+0.0007	+0.0021
Leu	+0.0036	-0.0042	+0.0216	-0.0048	+0.0018
Tyr	+0.0025	+0.0023	+0.0115	-0.0002	-0.0037
Asp	+0.0025	+0.0026	+0.0103	-0.0011	-0.0018
Lys	-0.0212	-0.0089	-0.0341	-0.0141	-0.0277
Gly	-0.0016	-0.0087	-0.0075	+0.0017	+0.0082
Glu	-0.0096	-0.0162	-0.0035	-0.0123	-0.0065
Ala	+0.0044	+0.0053	-0.0092	+0.0049	+0.0168
Pro	-0.0022	-0.0087	+0.0005	-0.0042	+0.0037

†The average number of gained (lost) residues per one amino acid substitution per site in the course of long-term evolution estimated for the two sets of proteins using PAML or EMAPI.
‡The number of gained (lost) residues of an amino acid per one amino acid substitution per site.

1. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13,** 555-556 (1997).
2. Brooks, D. J., Fresco, J. R. & Singh, M. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics* **20**, 2251 - 2257 (2004).

3. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol*., **1**, Research 8 (2001).
4. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol*. **13**, 1512-1517 (2003).
5. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci*. **8**, 275-282 (1992).

Supplementary Table 3. This table compares the pattern in amino acid gain and loss in recent evolution with the evidence on the order in which amino acids have been recruited into the genetic code. The rank of an amino acid suggested by the rate of its recent gain or loss (Fig. 1 in the main text) is in a good agreement with the consensus order of recruitment of amino acids into the genetic code, defined as a weighted average of the orders suggested by 60 empirical and theoretical criteria[1]. For five strong gainers and four strong losers, Spearman's correlation coefficient $r_s = 0.07$ ($P = 0.024$). Among the most credible empirical criteria are the results of experiments imitating conditions of prebiotic organic synthesis with electric discharge (spark) in a reducing atmosphere, pioneered by Miller[2,3]. Another important source of information on probable abiogenic amino acids are carbonaceous chondrites, a distinct class of meteorites containing a broad repertoire of organic compounds. Although caution is due in the interpretation of meteorite data because of the likelihood of terrestrial contamination[4,5], the amino acids in the Murchison meteorite are widely believed to be extra-terrestrial and abiogenic given the presence of several amino acids that do not occur in life forms on earth, the near equal amounts of enantiomers, and the unusual isotope ratios[6-8]. The data on amino acid synthesis in spark experiments and the amino acid content of the Murchison meteorite were compared separately with the ranks of amino acids suggested by the rates of their recent gain or loss. Amino acids with lowest rank (four strong losers) tend to be abundant in spark experiments, whereas five strong gainers (except Ser) are absent ($P = 0.04$, Fisher's exact test). Even a stronger pattern is observed for Murchison meteorite ($P = 0.008$).

Supplementary Table 3 **Recruitment of amino acids into the genetic code**

Amino acid[@]	Rank, according to the rate of recent gain or loss	Consensus order of recruitment into the genetic code	Abundance in spark experiments[#]	Abundance in Murchison meteorite[#]
Pro	1	5	+	++
Ala	2	2	+++	++
Glu	3	7	+	++
Gly	4	1	+++	+++
Lys	5	15	-	-
Asp	6	3	+	+
Tyr	7	18	-	-
Leu	8	8	+	+
Trp	9	20	-	-
Gln	10	11	-	-
Arg	11	10	-	-
Val	12	4	+	++
Ile	13	12	+	+
Thr	14	9	+	-
Asn	15	13	-	-
Phe	16	17	-	-
Ser	17	6	+	-
His	18	14	-	-
Met	19	19	-	-
Cys	20	16	-	-

[@]Amino acids are listed in the order of the increasing values of D (the opposite order is used in Fig. 1), i. e. in the order in which they were recruited into the genetic code, as suggested by the pattern in their ongoing gain and loss.
[#]+++: present in abundance, ++: present in moderate abundance, +: present, -: absent.

1. Trifonov, E. N. The triplet code from first principles. *J. Biomol. Struct. & Dyn.* **22**, 1-11 (2004).
2. Miller, S. L. A production of amino acids under possible primitive Earth conditions. *Science* **117,** 528-529 (1953).
3. Miller, S. L. Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harbor Symp. Quant. Biol*. **52**, 17–27 (1987).
4. Kvenvolden, K. A. Criteria for distinguishing biogenic and abiogenic amino acids--preliminary considerations. *Space Life Sci*. **4**, 60-68 (1973).
5. Glavin, D. P., Bada, J. L., Brinton, K. L. & McDonald, G. D. Amino acids in the Martian meteorite Nakhla. *Proc Natl Acad Sci USA* **96**, 8835-8838 (1999).

6. Kvenvolden, K., Lawless, J., Pering, K., Peterson, E., Flores, J., Ponnamperuma, C., Kaplan, I. R. & Moore, C. Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature* **228**, 923-926 (1970).
7. Cronin, J. R. & Pizzarello, S. Amino acids in meteorites. *Adv. Space Res*. **3**, 5-18 (1983).
8. Epstein, S., Krishnamurthy, R. V., Cronin, J. R., Pizzarello, S. & Yuen, G. U. Unusual stable isotope ratios in amino acid and carboxylic acid extracts from the Murchison meteorite. *Nature* **326**, 477-479 (1987).

Supplementary Methods. Correction for possible multiple substitutions at an amino acid site was introduced as follows. All the 15 pairs of sister genomes used in the present analysis are very close (Table 1). Thus, we assume that, at a site, no more than one substitution occurred after the divergence of sister species from their last common ancestor (CA). However, some outgroup species are distant from the sisters (Table 1). At a site, substitutions on the path connecting CA with the outgroup can either make identification of the CA amino acid impossible, or can lead to its misidentification. We used the following procedure to correct for the impact of such substitutions. Below, two subscripts denote amino acids present, at a given site, in the two sister species, the superscript denotes the amino acid in their CA, and the amino acid in the outgroup is shown in parentheses.

Let N^i_{ij} be the true number of sites with amino acids *i* and *j* in the two sister species and amino acid *i* in their CA. This number represents the total flux of $i > j$ substitutions. An estimate of N^i_{ij} has to be constructed on the basis of the amino acid observed in the outgroup:

$$N^i_{ij} = N^i_{ij}(i) + N^i_{ij}(j) + N^i_{ij}(x)$$

where $N^i_{ij}(i)$, $N^i_{ij}(j)$, and $N^i_{ij}(x)$ are numbers of sites where the outgroup carries *i* (CA state is identified correctly), *j* (CA state is misidentified), and some amino acid *x* different from both *i* and *j* (CA state remains unknown), respectively.

In order to estimate $N^i_{ij}(i)$, we note that

$$N^i_{ij}(i) = N_{ij}(i) - N^j_{ij}(i)$$

where $N_{ij}(i)$ is the number of all sites with amino acids *i* and *j* in the two sister species and amino acid *i* in the outgroup. This number is known, and $N^j_{ij}(i)$ can be inferred. The probability that the sister species have amino acids *i* and *j* and the outgroup has a third amino acid is $N_{ij}(x)/n$, where n is the total number of sites. Under the assumption of detailed equilibrium, CA states *i* and *j* are equally probable. The probability that the outgroup would have neither *i* nor *j* given either *i* or *j* in CA can be estimated as $(N_{jj}(x)/n_j + N_{ii}(x)/n_i)$, where n_i and n_j are the total numbers of sites occupied by amino acids *i* and j, respectively (these numbers are the same for all the species involved). The probability that the outgroup carries amino acid *i* given *j* in CA can be estimated as $N_{jj}(i)/n_j$. Therefore,

$$\mathrm{N}^{j}_{ij}(i) \approx N_{ij}(x)\frac{N_{jj}(i)}{n_j} \Big/ \left(\frac{N_{jj}(x)}{n_j} + \frac{N_{ii}(x)}{n_i}\right).$$

Analogously,

$$\mathrm{N}^{i}_{ij}(j) \approx N_{ij}(x)\frac{N_{ii}(j)}{n_i} \Big/ \left(\frac{N_{jj}(x)}{n_j} + \frac{N_{ii}(x)}{n_i}\right)$$

and

$$\mathrm{N}_{ij}^{i}(x) \approx N_{ij}(x)\frac{N_{ii}(x)}{n_i} \Big/ \left(\frac{N_{jj}(x)}{n_j} + \frac{N_{ii}(x)}{n_i}\right).$$

Thus, the flux of *i*->*j* substitutions, N^{i}_{ij}, is estimated as the sum of the above three terms.

Obviously, this estimate is conservative for our purposes, *i. e.*, it tends to underestimate the asymmetry of fluxes of reciprocal amino acid substitutions. Indeed, the detailed equilibrium was assumed. If, however, *i* is a gainer and *j* is a loser, the outgroup amino acid will be different from *i* in the CA less often than from *j* in the CA, *i. e*. the substitutions which remove a gainer will be recorded with a higher probability than substitutions which remove a loser.