doi 10.18699/bgrs2024-1.1-10

Связь информационной емкости и таксономии геномов оспы

Сенашова М.[1]*, Садовский М.[1, 2, 3]

[1] *Институт вычислительного моделирования СО РАН, Красноярск, Россия*

[2] *ФСНКЦ ФМБА России, Красноярск, Россия*

[3] *Сибирский федеральный университет, Красноярск, Россия*

* *msen@icm.krasn.ru*

Ключевые слова: нуклеотидная последовательность; частота; условная энтропия; главные компоненты; кластер

Мотивация и цель: Анализ статистических свойств нуклеотидных последовательностей является основной задачей биофизики, биоинформатики и молекулярной биологи. Множество работ было посвящено различным аспектам этого направления исследований (см., например, [1–4]). В данной работе рассмотрена связь между информационной емкостью частотных словарей нуклеотидных последовательностей толщиной от 2 до 20 символов и биологическими свойствами этих последовательностей. Каждой нуклеотидной последовательности ставится в соответствие точка в 18-мерном пространстве. Координатами точки являются условные энтропии, вычисленные для данной последовательности. Проанализирован состав кластеров, образованных проекциями точек из 18-мерного пространства в пространство первых трех главных компонент.
Методы и алгоритмы: Мы будем использовать частотные словари – совокупность всех фрагментов нуклеотидной последовательности фиксированной длины, встречающихся в ней, с указанием их частоты [5–7]. Введем основные понятия. Рассмотрим нуклеотидную последовательность длины N, состоящую из символов *A, C, G, T*. Любую последовательность символов длины q будем называть словом ω. Пусть n_ω – количество копий этого слова в нуклеотидной последовательности; тогда пару $\langle \omega,\ n_\omega \rangle$ будем называть (конечным) словарем *Wq* толщины *q*. Заменив количество копий частотой $f_\omega = \dfrac{n_\omega}{N}$, получаем частотный словарь *Wq* той же толщины. В работе [8] была изучена информационная емкость символьных последовательностей как условная энтропия частотных словарей толщины *q*, вычисленная по частотным словарям меньшей толщины. Мы используем этот подход для вычисления информационной емкости генетических последовательностей для разной толщины словаря. Информационная емкость или условная энтропия словаря толщины q может быть вычислена по формуле: $\bar{S} = 2S_{q-1} - S_q - S_{q-2}$ либо $\bar{S} = 2S_1 - S_2$ для $q = 2$, где $S_q = -\sum_{f_\omega} f_\omega \ln f_\omega$ – абсолютная энтропия частотного словаря толщины *q*. В нашем случае $q \le 20$. То есть для каждого генома мы получаем 18 значений условной энтропии.
Результаты: Были рассмотрены 70 геномов, относящихся к семейству Poxviridae, размещенных в GenBank. Для этих геномов были вычислены условные энтропии

для словарей толщиныойот 1 до 20. Набор условных энтропий для каждого генома рассматриваем как точку в 18-мерном пространстве. Далее в программе VidaExpert (http://bioinfo-out.curie.fr/projects/vidaexpert/) проецируем полученное множество точек в пространство первых трех главных компонент. Были проанализированы кластеры, состоящие из точек, чьи значения условной энтропии близки в метрике евклидова пространства. Были обнаружены следующие кластеры: оспа насекомых (AF063866 *Melanoplus sanguinipes entomopoxvirus O isolate Tucson,* AF250284 *Amsacta moorei entomopoxvirus,* AP013055 *Anomala cuprea entomopoxvirus DNA,* HF679131 *Adoxophyes honmai enomopoxvirus L,* HF679132 *Choristoneura biennis entomopoxvirus L,* HF679133 *Choristoneura rosaceana entomopoxvirus L,* HF679134 *Mythimna separata entomopoxvirus L*). Геном KR095315 *Diachasmimorpha longicaudata entomopoxvirus* не попал в этот кластер, так как *GC*-состав этого генома равен 0.3, а *GC*-составы остальных геномов оспы насекомых имеют значения около 0.2. Оспа птиц (AF198100 *Fowlpox virus,* AY318871 *Canarypox virus strain ATCC VR-111,* KJ801920 *Pigeonpox virus isolate FeP2,* KJ859677 *Penguinpox virus isolate PSan92,* MF678796 *Flamingopox virus FGPVKD09,* MK903864 *Magpiepox virus,* MT799800 *Cheloniid poxvirus 1,* MW365933 *Albatrosspox virus strain SAN97-0665NZ,* MW485973 *Magpiepox virus 2 isolate 62-11-06-2000-ANU*). В этот же кластер попал геном оспы черепахи MT799800 *Cheloniid poxvirus 1*. Кластер, в состав которого входят геномы, относящиеся к роду *Orthopoxvirus* (AF482758 *Cowpox virus strain Brighton Red,* AY009089 *Camelpox virus CMS,* DQ437594 *Taterapox virus strain Dahomey 1968,* HM172544 *Monkeypox virus strain Zaire 1979-005,* KP143769 *Raccoonpox virus,* KU749310 *Skunkpox virus strain WA,* KU749311 *Volepox virus strain CA,* L22579 *Variola major virus strain Bangladesh-1975,* MH607141 *Akhmeta virus isolate Akhmeta 2013-88,* MH816996 *Orthopoxvirus Abatino,* MN240300 *Borealpox virus,* OP526861 Monkeypox virus isolate MPXV USA 2022 FL0019, X69198 *Variola virus DNA,* Y16780 *Variola minor virus*). Кластер, в который входят геномы вирусов оспы парнокопытных (AF325528 *Lumpy skin disease virus NI-2490 isolate Neethling 2490,* AF410153 *Swinepox virus isolate 17077-99,* AY077832 *Sheeppox virus 10700-99 strain TU-V02127,* AY077835 *Goatpox virus Pellor,* AY689436 *Deerpox virus W-848-83,* AY689437 *Deerpox virus W-1170-84,* MF966153 *White-tailed deer poxvirus isolate OV179,* MG751778 *Moosepox virus GoldyGopher14*) и геномы вирусов AJ293568 *Yaba-like disease virus YLDV,* EF420156 *Tanapox virus isolate TPV-Kenya,* HQ849551 *Yoka poxvirus strain DakArB 4268*. Кластер, в который входят геномы оспы подсемейства *Chordopoxvirinae* разных родов (AF170722 *Rabbit fibroma virus,* AF170726 *Myxoma virus strain Lausanne,* KT159937 *Salmon gill poxvirus,* KU980965 *Pteropox virus strain Australia,* MH427217 *Sea otter poxvirus strain ELK,* MN653921 *Cetacean poxvirus 1 strain CePV-TA,* MT712273 *Teiidae poxvirus 1 isolate 1642 19*). Кластер, включающий геномы оспы летучих мышей, дикобраза и мышей (HQ647181 *Cotia virus SPAn232,* KY747497 *Eptesipox virus strain Washington,* MK860688 *Hypsugopox virus strain 251170-23 2017,* MN692191 *Brazilian porcupinepox virus 1 strain UFU USP001*). Кластер, содержащий геномы оспы подсемейства *Chordopoxvirinae* с *GC*-составом больше 0.6 (AY386265 *Bovine papular stomatitis virus strain BV-AR02,* GQ329670 *Pseudocowpox virus strain VR634,* HE601899 *Squirrel poxvirus strain Red squirrel UK,* KM502564 *Parapoxvirus red deer HL953 strain HL953,* MN339351 *Equine molluscum contagiosum-like virus strain Tanzania 2016,* U60315 *Molluscum contagiosum virus subtype 1*). Пары близко

расположенных точек, соответствующие геномам оспы близкородственных видов (DQ356948 *Nile crocodilepox virus*, MG450915 *Saltwater crocodilepox virus subtype 1*), (MF467280 *Western grey kangaroopox virus strain Western Australia*, MF467281 *Eastern grey kangaroopox virus strain Sunshine Coast*) и (MF001304 *Murmansk poxvirus strain LEIV-11411*, MF001305 *NY 014 poxvirus strain 2013*). А также геномы оспы, точки, соответствующие которым, находятся достаточно удаленно от описанных выше кластеров (KR095315 *Diachasmimorpha longicaudata entomopoxvirus*, KY382358 *Seal parapoxvirus isolate AFK76s1*, LC613089 *Carp edema virus FTI2020 DNA*, KP728110 *Turkeypox virus strain TKPV-HU1124 2011*, MF503315 *Squirrelpox virus Berlin 2015*).

Приведенная выше кластеризация геномов оспы показывает, что наблюдается достаточно сильная зависимость между статистическими свойствами геномов оспы (в нашем случае условной энтропии) и таксономией носителей заболевания либо таксономией самого вируса. Особенно четко это видно на геномах оспы насекомых, оспы птиц и оспы парнокопытных. В то же время имеются кластеры, в которые попали геномы оспы одного рода *Orthopoxvirus* или одного подсемейства *Chordopoxvirinae*.

Финансирование: Базовый проект «Новые методы и технологии комплексного анализа сложных природных и антропогенных экосистем на основе современных средств моделирования и обработки данных, распределенных вычислений и цифрового мониторинга», FWES-2024-0014.

Relationship between information capacity and taxonomy of smallpox genomes

Senashova M.[1]*, Sadovsky M.[1, 2, 3]

[1] *Institute of Computational Modelling, SB RAS, Kasnoyarsk, Russia*

[2] *FSR&CC of FMBA of Russia, Krasnoyarsk, Russia*

[3] *Siberian Federal University, Krasnoyarsk, Russia*

** msen@icm.krasn.ru*

Key words: nucleotide sequence; frequency; specific entropy; principal components; cluster

Motivation and Aim: Analysis of the statistical properties of nucleotide sequences is a core task of biophysics, bioinformatics and molecular biology. A number of papers have recently in this area (see, e.g. [1–4]). This paper examines the relationship between the information capacity of frequency dictionaries of nucleotide sequences with a thickness of 2 to 20 characters and the biological properties of these sequences. Each nucleotide sequence is assigned a point in 18-dimensional space. The coordinates of a point are the conditional entropies calculated for a given sequence. The composition of clusters formed by projections of points from 18-dimensional space into the space of the first three principal components is analyzed.

Methods and Algorithms: We will use frequency dictionaries it is a set of all fragments of a nucleotide sequence of a fixed length found in it, accompanied with their frequency [5–7]. An idea of the information capacity [8] characterizes sequences was studied as the conditional entropy of frequency dictionaries of thickness q, calculated from frequency dictionaries of smaller thickness (shorter words). The information capacity or

conditional entropy of the thickness dictionary q is provided by the formulae $\bar{S} = 2S_{q-1} - S_q - S_{q-2}$ or $\bar{S} = 2S_1 - S_2$ in case of $q = 2$; these are the absolute entropy of the thickness q of frequency dictionary. In our case $S_q = -\sum_{f_\omega} f_\omega \ln f_\omega$. Thus each genome is converted into a set of 18 figures of conditional entropy.

Results: 70 genomes of the family Poxviridae, from GenBank, have been examined. For these genomes, conditional entropies were calculated for dictionaries of thickness ranging from 1 to 20. The set of conditional entropies for each genome is considered as a point in 18-dimensional space. Next, the freely distributed VidaExpert software (http://bioinfo-out.curie.fr/projects/vidaexpert/) mapping the genomes into a set of points in principal components. We analyzed clusters consisting of points whose values of conditional entropy are close in the metric of Euclidean space. The following clusters were found: insect pox (AF063866 *Melanoplus sanguinipes entomopoxvirus O isolate Tucson*, AF250284 *Amsacta moorei entomopoxvirus*, AP013055 *Anomala cuprea entomopoxvirus DNA*, HF679131 *Adoxophyes honmai enomopoxvirus L*, HF679132 *Choristoneura biennis opoxvirus L*, HF679133 *Choristoneura rosaceana entomopoxvirus L*, HF679134 *Mythimna separata entomopoxvirus L*). The KR095315 *Diachasmimorpha longicaudata entomopoxvirus* genome did not fall into this cluster, since the *GC*-content of this genome is 0.3, and the *GC*-contents of the remaining insect pox genomes is about 0.2. Bird pox (AF198100 *Fowlpox virus*, AY318871 *Canarypox virus strain ATCC VR-111*, KJ801920 *Pigeonpox virus isolate FeP2*, KJ859677 *Penguinpox virus isolate PSan92*, MF678796 *Flamingopox virus FGPVKD09*, MK903864 *Magpiepox virus*, MT799800 *Che loniid poxvirus 1*, MW365933 *Albatrosspox virus strain SAN97-0665NZ*, MW485973 *Magpiepox virus 2 isolate 62-11-06-2000-ANU*). The same cluster included the turtlepox genome MT799800 *Cheloniid poxvirus 1*. A cluster including genomes of the genus *Orthopoxvirus* genomes (AF482758 *Cowpox virus strain Brighton Red*, AY009089 *Camelpox virus CMS*, DQ437594 *Taterapox* strain virus *Dahomey 1968*, HM172544 *Monkeypox virus strain Zaire 1979-005*, KP143769 *Raccoonpox virus*, KU749310 *Skunkpox virus strain WA*, KU749311 *Volepox virus strain CA*, L22579 *Variola major virus strain Bangladesh-1975*, MH607141 *Akhmeta virus isolate Akhmeta 2013-88*, MH816996 *Orthopoxvirus Abatino*, MN240300 *Borealpox virus*, OP526861 *Monkeypox virus isolate MPXV USA 2022 FL0019*, X69198 *Variola virus DNA*, *Y16780 Variola minor virus*). A cluster comprising the genomes of artiodactyl pox viruses (AF325528 *Lumpy skin disease virus NI-2490 isolate Neethling 2490*, AF410153 *Swinepox virus isolate 17077-99*, AY077832 *Sheeppox virus 10700-99 strain TU-V02127*, AY077835 *Goatpox virus Pellor*, AY689436 *Deerpox virus W -848-83*, AY689437 *Deerpox virus W-1170-84*, MF966153 *White-tailed deer poxvirus isolate OV179*, MG751778 *Moosepox virus GoldyGopher14*) and virus genomes AJ293568 *Yaba-like disease virus YLDV*, EF420156 *Tanapox virus isolate TPV-Kenya*, HQ849551 *Yoka poxvirus strain DakArB 4268*. Cluster comprising smallpox genomes of the subfamily *Chordopoxvirinae* of different genera (AF170722 *Rabbit fibroma virus*, AF170726 *Myxoma virus strain Lausanne*, KT159937 *Salmon gill poxvirus*, KU980965 *Pteropox virus strain Australia*, MH427217 *Sea otter poxvirus strain ELK*, MN653921 *Cetacean poxvirus 1 strain CePV-TA*, MT712273 *Teiidae poxvirus 1 isolate 1642 19*). A cluster comprising the smallpox genomes of bats, porcupines and mice (HQ647181 *Cotia virus SPAn232*, KY747497 *Eptesipox virus strain Washington*, MK860688 *Hypsugopox virus strain*

251170-23 2017, MN692191 *Brazilian porcupinepox virus 1 strain UFU USP001*). Cluster comprising smallpox genomes of the subfamily *Chordopoxvirinae* with *GC*-content greater than 0.6 (AY386265 *Bovine papular stomatitis virus strain BV-AR02*, GQ329670 *Pseudocowpox virus strain VR634*, HE601899 *Squirrel poxvirus strain Red squirrel UK*, KM502564 *Parapoxvirus red deer 3 strain HL953*, MN339351 *Equine molluscum contagiosum-like virus strain Tanzania 2016*, U60315 *Molluscum contagiosum virus subtype 1*). Pairs of closely located points corresponding to the smallpox genomes of closely related species: (DQ356948 *Nile crocodilepox virus*, MG450915 *Saltwater crocodilepox virus subtype 1*), (MF467280 *Western gray kangaroopox virus strain Western Australia*, MF467281 *Eastern gray kangaroopox strain virus Sunshine Coast*) and (MF001304 Murmansk poxvirus strain *LEIV-11411*, MF001305 *NY 014 poxvirus strain 2013*), as well as smallpox genomes, which corresponding points are located quite far from the clusters described above (KR095315 *Diachasmimorpha longicaudata entomopoxvirus*, KY382358 *Seal parapoxvirus isolate AFK76s1*, LC613089 *Carp edema virus FTI2020 DNA*, KP728110 *Turkeypox virus strain TKPV-HU1124 2 011*, MF503315 *Squirrelpox virus Berlin 2015*).

The above clustering of smallpox genomes shows that there is a fairly strong dependence between the statistical properties of smallpox genomes (in terms of conditional entropy) and the taxonomy of disease hosts, or the taxonomy of the virus itself. This is clearly are the genomes of insect pox, bird pox and artiodactyl pox. Simultaneously, there are clusters that contain smallpox genomes of one genus *Orthopoxvirus* or one subfamily *Chordopoxvirinae*.

Funding: Basic project “New methods and technologies for complex analysis of complex natural and anthropogenic ecosystems based on modern modeling and data processing tools, distributed computing and digital monitoring”, FWES-2024-0014.

Список литературы/References

1. Orlov Y.L., Boekhorst R.T., Abnizova I.A. Statistical measures of the structure of genomic sequences: Entropy, complexity, and position information. *J Bioinform Comput Biol.* 2006;4:523-536. doi 10.1142/S0219720006001801
2. Sabatti C., Rohlin L., Lange K., Liao J.C. Vocabulon: A dictionary model approach for reconstruction and localization of transcription factor binding sites. *Bioinformatics*. 2005;21:922-931. doi 10.1093/bioinformatics/bti083
3. Li W., Liu Y., Huang H.C. et al. Dynamical systems for discovering protein complexes and functional modules from biological networks. *IEEE/ACM Trans Comput Biol Bioinform*. 2007;4(2):233-250. doi 10.1109/TCBB.2007.070210
4. Гельфанд М.С. Апология биоинформатики. *Биофизика*. 2005;50(4):752-766
[Gelfand M.S. Apologia of bioinformatics. *Biophysics*. 2005;50(4):663-676 (in Russian)]
5. Bernaola-Galván P. et al. Study of statistical correlations in DNA sequences. *Gene*. 2002;300(1-2):105-115. doi 10.1016/s0378-1119(02)01037-5
6. Gorban A.N. et al. A new approach to the study of statistical properties of genetic sequences. *Biophysics*.1993;38(5):783-787
7. Горбань А.Н., Попова Т.Г., Садовский М.Г. Избыточность генетических текстов и мозаичная структура генома. *Молекулярная биология*. 1994;28(2):313-322
[Gorban A.N., Popova T.G., Sadovsky M.G. Redundancy of genetic sequences and mosaic structure of a genome. *Molekuliarnaia Biologiia*. 1994;28(2):313-322 (in Russian)]
8. Sadovsky M.G. Genomes and Information. *Biophysics*. 2009;54(4):419-422