УДК 004.94 doi: 10.15827/0236-235X.151.409-416 2025. Т. 38. № 3. С. 409–416

Многоуровневая оценка точности средств прогнозирования функциональной структуры генетических последовательностей

В.И. Аржаев [1], А.В. Скворцов [1]✉

[1] НИИ «Центрпрограммсистем», г. Тверь, 170024, Россия

Ссылка для цитирования
Аржаев В.И., Скворцов А.В. Многоуровневая оценка точности средств прогнозирования функциональной структуры генетических последовательностей // Программные продукты и системы. 2025. Т. 38. № 3. С. 409–416. doi: 10.15827/0236-235X.151.409-416
Информация о статье
Группа специальностей ВАК: 2.3.1
Поступила в редакцию: 22.01.2025 После доработки: 09.03.2025 Принята к публикации: 16.04.2025

Аннотация. В статье рассматриваются некоторые методики количественной оценки качества прогнозирования функциональной структуры генетических последовательностей доступными программами прогнозирования генов. Объектом исследования послужили алгоритмы и программные средства для прогнозирования функциональной структуры генетических последовательностей. Предметом являются количественные показатели точности прогнозирования функциональной структуры генетических последовательностей и алгоритмы их вычислений. Рассмотрено состояние разработки методов сравнения функциональной аннотации генетических последовательностей и самих методов прогнозирования функциональной структуры геномов. В результате отобраны количественные показатели схожести элементов функциональной аннотации нуклеотидных последовательностей на уровне отдельных нуклеотидов и экзон-интронной структуры генов, методики их расчета адаптированы для применения к оценке достоверности результата работы существующих программных средств прогнозирования генов на уровнях нуклеотидов и экзонов. На основе отобранных и уточненных методик сравнения эталонной и спрогнозированной функциональной структуры последовательностей разработано программное средство оценки качества прогнозирования кодирующих белки генов. Рассмотрено описание статической структуры разработанной программы и обобщенное описание алгоритма формирования статистических оценок качества прогнозирования генов по сравнению с эталонной функциональной аннотацией. В отличие от известных программных средств с открытым исходным кодом обеспечивается вычисление более информативных показателей точности прогнозирования генов по сравнению с базовыми метриками ложноположительных и ложноотрицательных результатов.
Ключевые слова: средства прогнозирования, функциональная структура, генетические последовательности, многоуровневая оценка, программное средство

Введение. В современной научно-инженерной практике в области обработки результатов молекулярно-генетических исследований биологических и биоподобных объектов с применением методов биоинформатики важное место занимают программные средства прогнозирования функциональной структуры генетических последовательностей без априорной информации (прогнозирование генов ab initio). В этом контексте под функциональной структурой генетических последовательностей понимается расположение генов, их экзон-интронная структура, наличие и расположение регуляторных элементов и других функционально значимых участков.

К настоящему времени разработаны различные методы решения задачи прогнозирования функциональной структуры генетических последовательностей, достаточно широко применяются реализующие их программные средства.

Наиболее простые, но все еще применяющиеся на практике подходы основаны на поиске открытых рамок считывания в нуклеотидных последовательностях. Их реализации доступны в открытых программных пакетах EMBOSS, где эту функцию выполняет программа getorf [1] и ПО Ugene, в состав которого входит компонент find-orfs [2]. Более сложные подходы используют статистические методики вычисления потенциала кодирования белка фрагментом генетической последовательности (методика расчета показателя кодирующего потенциала, опубликованная в [3] и усовершенствованная в [4]). Реализации этих методик также доступны в пакете EMBOSS (программа tcode) и в интернет-сервисе CPC2 [5]. Следующий шаг в развитии методов прогнозирования генов представляет программа prodigal [6], которая интегрирует расчет кодирующего потенциала фрагмента генетической последовательности в алгоритм динамического программирования, оптимизирующий формируемый для анализируемой последовательности набор генов по нескольким статистическим критериям.

Исторически и в направлении возрастания сложности располагаются подходы, использующие алгоритмы машинного обучения. До недавнего времени лучшие результаты демонстрировали программы, основанные на обучении скрытых марковских моделей, представляющих структуру последовательностей как смену скрытых состояний генов и регуляторных участков, в то время как наблюдаемыми состояниями являются символы последовательности. Среди программ, реализующих этот метод, широко используется Augustus [7]. К преимуществам относится сравнительная легкость адаптации к генетическим последовательностям новых организмов за счет обучения на экспериментально подтвержденных данных о структурных элементах последовательностей.

В последние годы стремительное развитие моделей искусственного интеллекта для обработки текстов на естественном языке способствовало появлению нового класса инструментов биоинформатики: геномных языковых моделей, одним из применений которых также является прогнозирование функциональной структуры ДНК-последовательностей. В работе [8] описывается модель прогнозирования функциональной структуры ДНК-последовательностей Nucleotide transformer, исследование [9] посвящено геномной языковой модели анализа процессов регуляции синтеза белков и их функций, в работе [10] рассмотрены общие вопросы структуры и обучения геномных языковых моделей.

Несмотря на большое количество доступных программных средств для прогнозирования генов их точность и применимость для использования с конкретными генетическими последовательностями в медико-биологических исследованиях остается неоднозначной и зависит от множества факторов, включая специфику организма источника анализируемых последовательностей, используемый алгоритм и степень его параметризации. В ряде исследований предлагались метрики оценки качества прогнозирования генов. Работа [11] описывает расчет традиционных метрик чувствительности и специфичности для нуклеотидного и экзонного уровня, в [12] вводятся производные показатели, исследование [13] вводит метрики экзон-интронного уровня, статья [14] посвящена оценке современных средств прогнозирования генов доступными методами. Однако рассмотренные работы не позволяют всесторонне охарактеризовать эффективность и применимость инструментов прогнозирования на разных уровнях организации генетической информации – от нуклеотидного до функционального. Это затрудняет выбор оптимального ПО для конкретной задачи и снижает надежность получаемых результатов.

Метод

В работах [11–14] предлагаются метрики точности прогнозирования, которые могут быть адаптированы для оценки достоверности результатов предсказания структуры генетических последовательностей: отдельных нуклеотидов, экзонов, генов, белковых продуктов.

Пусть даже биологический смысл функциональной аннотации ДНК-последовательностей очевиден, результат ее выполнения и, следовательно, процесс сравнения результата с эталонным для формирования оценки качества работы программного средства, выполнявшего эту аннотацию, может быть легко освобожден от биологического смысла, и задача будет ограничена обработкой разметки текстовых строк с выделением фрагментов нескольких типов.

На уровне отдельных нуклеотидов генетическая последовательность может быть представлена строкой символов, где кодирующие белки гены, а точнее, экзоны – фрагменты генов, являются участками этой строки ненулевой длины, возможно, пересекающимися.

Для оценки качества прогнозирования кодирующих белки генов на уровне нуклеотидов, как и на всех последующих уровнях, используется эталонная аннотация – экспериментально подтвержденные данные о кодирующих белки генах в последовательности, которая используется для тестирования конкретного программного средства. Каждому символу строки последовательности присваиваются признаки принадлежности к кодирующему фрагменту в эталонной аннотации и в прогнозной (полученной в результате работы тестируемого программного средства):

– *AP* (actual positive) – является частью кодирующего региона в эталонной аннотации;

– *AN* (actual negative) – не является частью кодирующего региона в эталонной аннотации;

– PP (predicted positive) – является частью кодирующего региона в прогнозной аннотации;

– *PN* (predicted negative) – не является частью кодирующего региона в прогнозной аннотации.

Соотношение этих признаков определяет принадлежность символа в конкретной позиции p (N_p) следующим множествам (рис. 1):

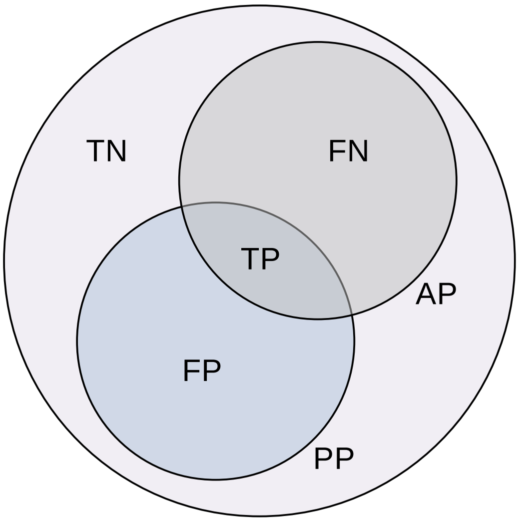

Рис. 1. Классификация элементов функциональной аннотации на нуклеотидном уровне

Fig. 1. Classification of nucleotide-level functional annotation elements

– *TP* (true positive) – истинноположительный результат прогнозирования принадлежности символа кодирующему региону при $N_p \in AP \cap PP$;

– *TN* (true negative) – истинноотрицательный результат при $N_p \in AN \cap PN$;

– *FP* (false positive) – ложноположительный результат при $N_p \in AN \cap PP$;

– *FN* (false negative) – ложноотрицательный результат при $N_p \in AP \cap PN$.

Приведем пример классификации элементов функциональной аннотации для фрагмента последовательности, содержащего перекрывающиеся кодирующие регионы в обеих аннотациях (рис. 2). Признак (*TP*, *TN*, *FP*, *FN*) присваивается каждому символу последовательности, попадающему в соответствующий диапазон на рисунке.

Наиболее распространенными метриками точности при использовании классификации элементов аннотации, аналогичной приведенной выше, являются чувствительность (*Sn*) и специфичность (*Sp*):

$$Sn_n = \frac{TP_n}{TP_n + FN_n} = \frac{TP_n}{AP_n},\ Sp_n = \frac{TP_n}{TP_n + FP_n} = \frac{TP_n}{PP_n}.$$

Здесь и далее нижний индекс *n* обозначает принадлежность параметров и коэффициентов нуклеотидному уровню оценки качества прогнозирования кодирующих белки областей генетической последовательности. Чувствительность характеризует способность тестируемого средства прогнозирования генов детектировать кодирующие регионы последовательности и является условной вероятностью $P(F(x) = c \mid x = c)$ события, состоящего в том, что в результате прогнозирования нуклеотид определен как принадлежащий кодирующему региону при условии, что он принадлежит кодирующему региону в эталонной аннотации. В то же время специфичность связана со способностью избегать ложного детектирования некодирующих областей как кодирующих и является условной вероятностью $P(x = c \mid F(x) = c)$ события, состоящего в том, что нуклеотид принадлежит кодирующему региону в эталонной аннотации при условии, что метод прогнозирования отнес его к кодирующему региону. Здесь и далее $F(x)$ – результат классификации нуклеотида в позиции *x* на кодирующий *c* или некодирующий *n*. Поскольку ни одна из этих двух характеристик не может быть использована для оценки метода или средства прогнозирования без другой, на практике часто используют различные составные меры. В простейшем случае это среднее значение чувствительности и специфичности *Ac* (accuracy):

$$Ac_n = \frac{(Sn_n + Sp_n)}{2}.$$

В работе [4] дополнительно вводится единый скалярный оценочный коэффициент – коэффициент корреляции (*Correlation Coefficient*, CC):

$$CC_n = \frac{TP_nTN_n - FN_nFP_n}{\sqrt{(TP_n + FN^n)(TN^n + FP_n) \times (TP_n + FP_n)(TN_n + FN_n)}} = \frac{TP_nTN_n - FN_nFP_n}{\sqrt{PP_nPN_nAP_nAN_n}}.$$

Однако у данного показателя есть очевидный недостаток – он не определен в случае, если *PP*, *PN*, *AP* или *AN* равны 0. Такое условие включает предельные случаи, когда эталонная или прогнозная аннотация не содержит кодирующих областей (последовательность полностью некодирующая) или не содержат интронов и междугенных регионов (вся последовательность кодирующая).

Наиболее простой используемой метрикой, определенной для любых данных, является простой коэффициент соответствия (*Simple Matching Coefficient*, SMC):

$$SMC_n = \frac{TP_n + TN_n}{TP_n + FN_n + FP_n + TN_n}.$$

Данная метрика также имеет ограничения, поскольку может принимать высокие значения

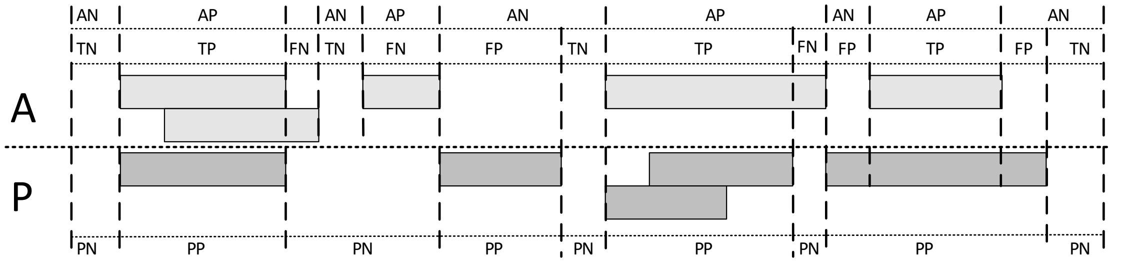

Рис. 2. Пример классификации на нуклеотидном уровне при наличии перекрывающихся элементов в эталонной и в прогнозной аннотациях

Fig. 2. Example of nucleotide-level classification with overlapping elements in reference and predicted annotations

при низкой чувствительности или специфичности, но высоком значении второй величины.

Наиболее информативной оценкой из рассмотренных является предложенный в той же работе коэффициент средней условной вероятности для всех классов (*Average Conditional Probability*, ACP). Кроме описанных условных вероятностей Sn и Sp, он зависит от $P(F(x) = n \mid x = n)$:

$$P_n^{PN} = \frac{TN_n}{TN_n + FP_n},$$

и $P(x = n \mid F(x) = n)$:

$$P_n^{AN} = \frac{TN_n}{TN_n + FN_n}.$$

Коэффициент ACP является средним значением четырех указанных условных вероятностей и в общем случае, когда определены все четыре, вычисляется следующим образом:

$$ACP_n = \frac{1}{4}\left(\begin{array}{l}\dfrac{TP_n}{TP_n + FN_n} + \dfrac{TP_n}{TP_n + FP_n} + \\ + \dfrac{TN_n}{TN_n + FP_n} + \dfrac{TN_n}{TN_n + FN_n}\end{array}\right) =$$

$$= \frac{1}{4}\left(\frac{TP_n}{PP_n} + \frac{TP_n}{AP_n} + \frac{TN_n}{PN_n} + \frac{TN_n}{AN_n}\right).$$

Для перевода ACP к диапазону значений, близкому к CC, используется преобразование, дающее значение приближенной корреляции (*Approximate Correlation*, AC):

$$AC_n = \left(ACP - \frac{1}{2}\right)2.$$

Точность прогнозирования генов может быть также оценена на уровне экзон-интронной структуры. В данном случае признаки *TP*, *TN*, *FP*, *FN* следует назначать не отдельным символам строки последовательности, а ее фрагментам, которые соответствуют экзонам функциональной структуры. При этом применяются в целом аналогичные метрики, однако сложность состоит в том, что не существует единых критериев определения корректно распознанного экзона. Так, корректно спрогнозированным может считаться экзон при полном совпадении границ с эталонной аннотацией, при перекрытии с эталонным на количество нуклеотидов, больше заданного порогового значения, при совпадении хотя бы одной границы и т.д.

Как и на нуклеотидном уровне, экзоны $E\ (b, e)$, где b – начальная граница экзона в последовательности, e – конечная граница, обладают следующими признаками, характеризующими их присутствие в аннотации:

AE (actual exon) – экзон является частью кодирующего белок гена в эталонной аннотации;

PE (predicted exon) – экзон является частью кодирующего белок гена в спрогнозированной аннотации;

CE (correct exon) – корректный экзон при $E\ (b, e) \in AE \cap PE$.

Признаки *AN* и *PN*, использованные на нуклеотидном уровне, не имеют смысла на экзонном уровне, так как вне кодирующих белки регионов последовательности не существует экзонов.

Здесь и далее нижний индекс e обозначает принадлежность параметров и коэффициентов экзонному уровню оценки качества прогнозирования кодирующих белки областей генетической последовательности. Применение приведенной выше модели признаков предоставляет возможность использовать на данном уровне описанные на предыдущем чувствительность и специфичность:

$$Sn_e = \frac{CE_e}{AE_e},\ \ Sp_e = \frac{CE_e}{PE_e}.$$

В данном случае появляется возможность использовать дополнительные метрики точности прогнозирования, которые основываются на том, что часть некорректно распознанных экзонов (*PE*) частично перекрывается с ошибочно не определенными (*AE*), в то время как некоторые спрогнозированные (*PE*) целиком располагаются вне кодирующих регионов в соответствии с эталонной аннотацией и являются неверными *W* (wrong). Часть экзонов – экзоны из эталонной аннотации (*AE*) – целиком располагаются в некодирующем регионе прогнозной аннотации и являются пропущенными *M* (missing). В связи с этим оказывается полезным ввести метрики доли пропущенных *MR* (miss rate) и доли неверных экзонов *WR* (wrong rate) [13]:

$$MR_e = \frac{M_e}{AE_e},\ WR_e = \frac{W_e}{PE_e}.$$

На рисунке 3 приведен тот же набор генов эталонной и прогнозной аннотаций, что и на рисунке 2, с разметкой признаков для экзонного уровня.

Оценка качества прогнозирования на уровне экзонов имеет более высокое значение по сравнению с нуклеотидным уровнем, поскольку любые ошибки в определении набора и границ экзонов ведут к неверному результату прогнозирования последовательностей кодирующих геном белков.

В качестве примера существующих программных средств с открытым исходным кодом для решения задачи оценки качества прогнозирования генов можно привести средство GFFCompare из пакета GFF Tools [15]. Данная программа проводит сопоставление аннотаций на большом числе уровней: дельных нуклеотидов, экзонов, интронов, цепочек интронов, транскриптов, но выполняет только вычисление базовых показателей *TP*, *FP*, *FN*.

Результаты

Описанные методики реализованы в программном средстве оценки качества прогнозирования функциональной структуры генетических последовательностей. Построена диаграмма статической структуры данной программы в нотации диаграммы классов UML2 (http://www.swsys.ru/uploaded/image/2025-3/17.jpg).

Основу статической структуры программы составляют классы сущностных объектов:

– Statistics: базовая статистика (TP, FP, FN, чувствительность, специфичность);

– NucleotideStatistics: расширенная статистика для нуклеотидов (TN, корреляции);

– ExonStatistics: статистика для экзонов (пропущенные, неправильные);

– EvaluationResult: объединяет статистику всех уровней;

– FeatureTreeItem: элемент дерева функциональных элементов аннотаций последовательностей.

Класс управляющего объекта Evaluator осуществляет построение дерева функциональных элементов и расчет исходных данных для вычисления показателей качества прогнозирования.

Листинг содержит краткое описание алгоритма работы программы в виде псевдокода:

```
1. Основная программа:
  - Получить аргументы командной строки (-a для эталонного файла, -p для проверяемого файла)
  - Открыть файлы
  - Вызвать функцию evaluate()

2. Функция evaluate(annoFile, evaluateFile):
```

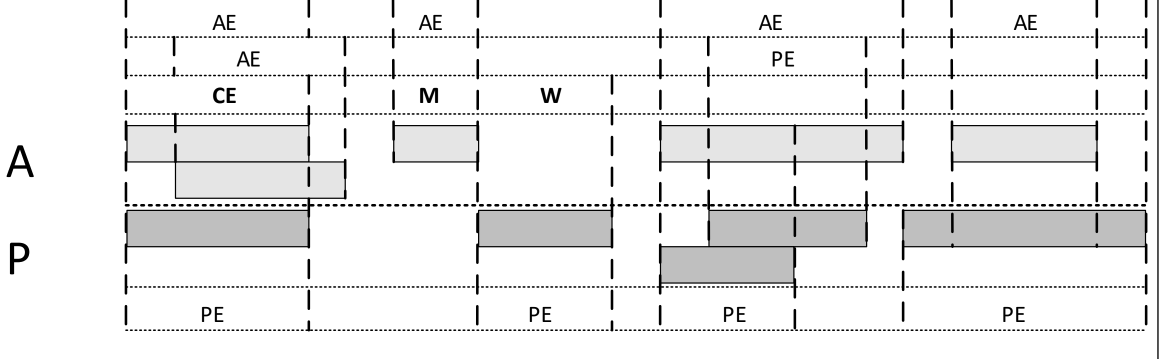

Рис. 3. Пример классификации идентифицированных функциональных элементов последовательности на экзонном уровне при наличии перекрывающихся элементов как в эталонной, так и в прогнозной аннотациях

Fig. 3. Classification example of identified functional sequence elements at an exon level with overlapping features in both reference and predicted annotations

Пример работы программы оценки качества прогнозирования генов для сравнения работы Ugene и Augustus

Example of gene prediction quality evaluation: Ugene and Augustus comparison

Организм	Последовательность	Программа прогнозирования	Уровень нуклеотидов			Уровень экзонов			
			Sen	Spec	AC	Sen	Spec	MR	WR
Escherichia coli	NC_000913	Ugene	0,987	0,618	0,554	0,728	0,135	0,011	0,523
		Augustus	0,964	0,997	0,966	0,834	0,892	0,062	0,004
Drosophila melanogaster	NC_004353	Ugene	0,799	0,221	0,341	0,004	0,001	0,249	0,837
		Augustus	0,566	0,871	0,691	0,327	0,570	0,464	0,164
	NC_024512	Ugene	0,676	0,227	0,28	0,01	0,003	0,318	0,756
		Augustus	0,914	0,911	0,902	0,773	0,828	0,111	0,075

```
    - Прочитать последовательности из файлов в формате GenBank
    - Для каждой цепи (+ и -):
      a. Построить дерево аннотаций (makeTree)
      b. Сравнить аннотации (объект класса Evaluator)
      c. Сохранить статистику
    - Объединить статистику для обеих цепей
    - Вывести и сохранить отчет

  3. Функция makeTree(sequence, strand):
    - Создать список генов
    - Для каждого элемента аннотации:
      - Если это ген: создать новый FeatureTreeItem
      - Если это CDS (кодирующая область): добавить экзоны к соответствующему гену
    - Удалить дубликаты экзонов
    - Вернуть список генов с экзонами

  4. Класс Evaluator:
    - Инициализация: принимает списки аннотированных и проверяемых генов
    - Сравнение:
      a. Уровень генов: подсчет TP, FP, FN
      b. Уровень экзонов: подсчет совпадений, пропущенных, неправильных
      c. Уровень нуклеотидов: точность предсказаний
    - Расчет статистики: чувствительность, специфичность, точность
```

В таблице приведены результаты применения программного средства оценки качества прогнозирования генов для сравнения результатов работы программ Ugene get_orfs и Augustus по прогнозированию генной структуры референсных генетических последовательностей двух модельных организмов.

Выводы

Разработка и применение многоуровневого подхода к оценке точности ab initio методов прогнозирования функциональной структуры генетических последовательностей представляет собой важную научно-практическую задачу.

Предложенные в работе показатели оценки точности работы программ прогнозирования генов и разработанное на их основе программное средство позволяют сравнивать различные алгоритмы объективно и детализировано на уровнях нуклеотидов и экзонов, что дает возможность подбирать подходящие средства ab initio прогнозирования генов для конкретных задач.

Список литературы

1. Rice P., Longden I., Bleasby A. EMBOSS: The European molecular biology open software suite. Trends in Genetics, 2000, vol. 16, no. 6, pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

2. Rose R., Golosova O., Tiunov A. et al. Flexible design of multiple metagenomics classification pipelines with UGENE. Bioinformatics, 2018, vol. 35, no. 11, pp. 1963–1965. doi: 10.1093/bioinformatics/bty901.

3. Fickett J.A., Tung C.S. Assessment of protein coding measures. Nucleic Acids Research, 1992, vol. 20, no. 24, pp. 6441–6450. doi: 10.1093/nar/20.24.6441.

4. Kang Y.-J., Yang D.-C., Kong L. et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research, 2017, vol. 45, no. W1, pp. W12–W16. doi: 10.1093/nar/gkx428.

5. Coding Potential Calculator 2. URL: https://cpc2.gao-lab.org/ (дата обращения: 21.01.2025).

6. Hyatt D., Chen G.-L., LoCascio P.F. et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 2010, vol. 11, art. 119. doi: 10.1186/1471-2105-11-119.

7. Lars G., Bruna T., Hoff K.J. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. Genome Research, 2024, vol. 34, no. 5, pp. 769–777. doi: 10.1101/2023.06.10.544449.

8. Dalla-Torre H., Gonzalez L., Mendoza-Revilla J. et al. Nucleotide transformer: Building and evaluating robust foundation models for human genomics. Nature Methods, 2025, vol. 22, pp. 287–297. doi: 10.1038/s41592-024-02523-z.

9. Hwang Y., Cornman A.L., Kellogg E.H. et al. Genomic language model predicts protein co-regulation and function. Nature Communications, 2024, no. 15, art. 2880. doi: 10.1038/s41467-024-46947-9.

10. Benegas G., Ye C., Albors C., Li J.C., Song U.S. Genomic language models: Opportunities and challenges. Trends in Genetics, 2025, vol. 41, no. 4, pp. 286–302.

11. Burset M., Guigo R. Evaluation of gene structure prediction programs. Genomics, 1996, vol. 34, no. 3, pp. 353–367. doi: 10.1006/geno.1996.0298.

12. Pavy N., Rombauts S., Dehais P. et al. Evaluation of gene prediction software using a genomic data set: Application to Arabidopsis thaliana sequences. Bioinformatics, 1999, vol. 15, no. 11, pp. 887–899. doi: 10.1093/bioinformatics/15.11.887.

13. Goodswen S.J., Kennedy P.J., Ellis J.T. Evaluating high-throughput Ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. PloS One, 2012, vol. 7, no. 11, art. e50609. doi: 10.1371/journal.pone.0050609.

14. Kirkland T.N., Beyhan S., Stajich J.E. Evaluation of different gene prediction tools in coccidioides immitis. J. Fungi, 2023, vol. 9, no. 11, art. 1094. doi: 10.3390/jof9111094.

15. Pertea G., Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research, 2020, no. 9, art. 304. doi: 10.12688/f1000research.23297.1.

Software & Systems doi: 10.15827/0236-235X.151.409-416 2025, 38(3), pp. 409–416

Multilevel accuracy assessment of prediction tools for a genetic sequence functional structure

Valentin I. Arzhaev [1], Andrey V. Skvortsov [1]✉

[1] Research Institute Centerprogramsystem, Tver, 170024, Russian Federation

For citation

Arzhaev, V.I., Skvortsov, A.V. (2025) ‘Multilevel accuracy assessment of prediction tools for a genetic sequence functional structure’, *Software & Systems*, 38(3), pp. 409–416 (in Russ.). doi: 10.15827/0236-235X.151.409-416

Article info

Received: 22.01.2025 After revision: 09.03.2025 Accepted: 16.04.2025

Abstract. The paper examines several methodologies for quantitative assessment of functional structure prediction quality in genetic sequences using available gene prediction tools. It focuses on algorithms and software tools for predicting functional structure in genetic sequences. The paper analyzes quantitative prediction accuracy parameters for functional structure in genetic sequences and their algorithmic implementations. The authors explored current development of methods for comparing functional annotation of genetic sequences, as well as methods for predicting a genome functional structure. As a result, they selected quantitative similarity metrics for functional annotation elements in nucleotide sequences. These metrics involve both nucleotide-level resolution and gene exon-intron structure. The calculation methods were adapted to assess the reliability of existing gene prediction software outputs. The methodology is applicable at both nucleotide and exon levels. Using the selected and refined methods for comparing reference and predicted functional sequence structures, the authors developed a software tool for assessing protein-coding gene prediction quality. The paper describes the static architecture of the developed program and a generalized algorithm for generating statistical quality metrics comparing gene predictions against reference annotations. Unlike existing open-source tools, the proposed solution calculates more informative gene prediction accuracy metrics that surpass basic false-positive and false-negative measures.

Keywords: prediction tools, functional structure, genetic sequence, multilevel accuracy, software tool

References

1. Rice, P., Longden, I., Bleasby, A. (2000) ‘EMBOSS: The European molecular biology open software suite’, *Trends in Genetics*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

2. Rose, R., Golosova, O., Tiunov, A. et al. (2025) ‘Flexible design of multiple metagenomics classification pipelines with UGENE’, *Bioinformatics*, 35(11), pp. 1963–1965. doi: 10.1093/bioinformatics/bty901.

3. Fickett, J.A., Tung, C.S. (1992). ‘Assessment of protein coding measures’, *Nucleic Acids Research*, 20(24), pp. 6441–6450. doi: 10.1093/nar/20.24.6441.

4. Kang, Y.-J., Yang, D.-C., Kong, L. et al. (2017) 'CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features', *Nucleic Acids Research*, 45(W1), pp. W12–W16. doi: 10.1093/nar/gkx428.

5. Coding Potential Calculator 2, available at: https://cpc2.gao-lab.org/ (accessed January 21, 2025) (in Russ.).

6. Hyatt, D., Chen, G.-L., LoCascio, P.F. et al. (2010) 'Prodigal: Prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11, art. 119. doi: 10.1186/1471-2105-11-119.

7. Lars, G., Bruna, T., Hoff, K.J. et al. (2024) 'BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA', *Genome Research*, 34(5), pp. 769–777. doi: 10.1101/2023.06.10.544449.

8. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. (2025) 'Nucleotide transformer: Building and evaluating robust foundation models for human genomics', *Nature Methods*, 22, pp. 287–297. doi: 10.1038/s41592-024-02523-z.

9. Hwang, Y., Cornman, A.L., Kellogg, E.H. et al. (2024) 'Genomic language model predicts protein co-regulation and function', *Nature Communications*, (15), art. 2880. doi: 10.1038/s41467-024-46947-9.

10. Benegas, G., Ye, C., Albors, C., Li, J.C., Song, U.S. (2025) 'Genomic language models: Opportunities and challenges', *Trends in Genetics*, 41(4), pp. 286–302.

11. Burset, M., Guigo, R. (1996) 'Evaluation of gene structure prediction programs', *Genomics*, 34(3), pp. 353–367. doi: 10.1006/geno.1996.0298.

12. Pavy, N., Rombauts, S., Dehais, P. et al. (1999) 'Evaluation of gene prediction software using a genomic data set: Application to Arabidopsis thaliana sequences', *Bioinformatics*, 15(11), pp. 887–899. doi: 10.1093/bioinformatics/15.11.887.

13. Goodswen, S.J., Kennedy, P.J., Ellis, J.T. (2012) 'Evaluating high-throughput Ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques', *PloS One*, 7(11), art. e50609. doi: 10.1371/journal.pone.0050609.

14. Kirkland, T.N., Beyhan, S., Stajich, J.E. (2023) 'Evaluation of different gene prediction tools in coccidioides immitis', *J. Fungi*, 9(11), art. 1094. doi: 10.3390/jof9111094.

15. Pertea, G., Pertea, M. (2020) 'GFF Utilities: GffRead and GffCompare', *F1000Research*, (9), art. 304. doi: 10.12688/f1000research.23297.1.

Авторы
Аржаев Валентин Иванович [1], к.т.н.,
заместитель генерального директора,
arzhaevvi@cps.tver.ru
Скворцов Андрей Викторович [1],
заместитель директора центра 6,
skvortsovav@cps.tver.ru

[1] НИИ «Центрпрограммсистем»,
г. Тверь, 170024, Россия

Authors
Valentin I. Arzhaev [1], Cand. of Sci. (Engineering),
Deputy Director General,
arzhaevvi@cps.tver.ru
Andrey V. Skvortsov [1],
Deputy Director of Center 6,
skvortsovav@cps.tver.ru

[1] Research Institute Centerprogramsystem,
Tver, 170024, Russian Federation