ParSeq Anwendungsbeschreibung

1 Welche Fragestellungen können mit ParSeq bearbeitet werden?

Das Programm ist in der Lage auf langen Zeichenketten nach dem Vorkommen von Strings zu suchen, die durch spezielle reguläre Ausdrücke beschrieben werden können. Für diese regulären Ausdrücke steht ein erweitertes Alphabet zur Verfügung, mit dem neben Zeichenfolgen auch biochemische Eigenschaften beschrieben werden können.

Biologischer Hintergrund:

Lange DNA-Sequenzen, etwa Bakteriengenome,Teile eukaryotischer Genome oder die Aminosäuresequenzen der Proteine können als Zeichenketten aufgefasst werden. Diese sollen nach dem Vorhandensein von charakteristischen oder funktionstragenden Bereichen durchsucht werden. (DNA: Promotoren, Proteinbindungsstellen. Aminosäuresequenz: Funktionelle Gruppen etc)
Solche Abschnitte sind oft durch erhaltene Sequenzen (Consensussequenzen) gekennzeichnet. Sie können einzeln oder als Muster (als Folge mehrerer Motive in definiertem Abstand zueinander) vorliegen. Das Programm ist in der Lage, sowohl nach vorgegebenen Sequenzen (mit und ohne zulässige Fehler), als auch nach einer Abfolge von mehreren Motiven zu suchen, dabei kann der Abstand der Motive variabel sein.
Wird die DNA in eine Aminosäuresequenz übersetzt, können die entstehenden Zeichenketten zudem nach Teilsequenzen durchsucht werden, die definierte biochemische Eigenschaften besitzen, soweit sich diese aus den sequenzbildenden Aminosäuren ableiten lassen (Berechnung unter Verwendung von Scalingtabellen). So kann auch auf Aminosäuresequenzen nach Bereichen bestimmter Funktion gesucht werden, sofern Bedingungen (Consensussequenzen, Bereiche bestimmter biochemischer Eigenschaften) beschrieben werden können, deren Erfüllung mit hoher Wahrscheinlichkeit an das Vorhandensein der biologischen Funktion gekoppelt ist. (Suche nach reaktiven Gruppen, Signalsequenzen etc)

2 Was wird als Eingabe benötigt?

Als Eingabe werden die zu durchsuchenden Sequenzen als Textfile und der reguläre Ausdruck, nach dem gesucht werden soll, benötigt.
Das Programm bietet die Möglichkeit Sequenzdateien (Testfiles), die über das Filesystem verfügbar sind, für eine Suchabfrage auszuwählen. Das Programm merkt sich automatisch den Pfad der letzten ausgewählten Datei. Der Benutzer kann unterscheiden zwischen DNA und Protein-Dateien. Je nachdem welche Art gewählt wurde, wird eine andere Abkürzungstabelle verwendet (s. Anhang) und die Ergebnisausgabe enthält bei Proteindateien zusätzlich eine Umrechnung auf die DNA Position. Die Suche bezieht sich immer auf alle ausgewählten Dateien. Es kann immer nur auf DNA oder nur auf Proteindateien gesucht werden.
Der Pfad zu den ausgewählten Dateien wird für weitere Abfragen gespeichert, außerdem werden die Zwischenergebnisse in einem Verzeichnisbaum abgelegt, auf den gefundenen Zwischenergebnissen kann anschließend weiter gesucht werden. Dadurch kann eine grobe Suche durchgeführt und anschließend schrittweise verfeinert werden.

3 Syntax der regulären Ausdrücke

Die zu suchenden Muster werden durch erweiterte reguläre Ausdrücke beschrieben. Als Grundlage wird die Syntax von existierenden Libraries (java.util.regex, regex++, Perl …) verwendet. Sie wird durch einige zusätzliche Möglichkeiten ergänzt. Einfachste Bausteine sind Zeichenketten. Teilausdrücke werden durch „@“ getrennt

Gesucht wird eine Zeichenkette die ...	**Beispiel**
...exakt der beschriebenen entspricht	(ABC)*
...beliebige Zeichen enthalten kann, die Länge des Strings ist durch minimale und maximale Länge charakterisiert	(X{min,max})*
...wahlweise eines der angegebenen Zeichen enthält,	(A|B|C)
...den angegebenen String so oft wie in Anzahl angegeben wiederholt umfasst	((ABC){Anzahl})

* hier können die runden Klammern auch weggelassen werden.

Beispiel:
(ABBA)@X{3,5}@(A|B|D)
Beschreibung:
Gesucht werden soll ein exakter String ABBA, gefolgt von einer 3 bis 5 Zeichen langen Zeichenkette beliebiger Zeichen, gefolgt von A oder B oder D.
Zusätzlich können für jeden Teilausdruck (bis zum nächsten @-Zeichen) Nebenbedingungen bestimmt werden.
Nebenbedingungen sind biochemische Eigenschaften oder Distanzen.
Die Nebenbedingungen haben definierte Namen und eine Argumentliste. Sie werden durch „/“ angehängt. Nach „/“ folgt eine Semikolon-Liste der Argumente.
(Liste der Namen vgl. Anhang)

Beispiel:
(ABC)@X{20}/hdp_kd (5,>,0)@
Beschreibung:
Gesucht werden soll ein exakter String ABC, gefolgt von einer 20 Zeichen langen Zeichenkette beliebiger Zeichen (Aminosäuren), wobei diese Zeichenkette einen Hydrophobizitätsscore nach Kyte/Doolittle größer 0, gemessen als Mittelwert in einem 5 Zeichen großen Fenster, haben soll.
(Die Fenstergröße über die gemittelt wird, ist zwischen 1 und 25 wählbar. Außer > sind < und = als Vergleichsoperatoren möglich.)

Beispiel:
(ABCDEFG)/hd(2)
Beschreibung:
Gesucht wird die angegebene Zeichenkette, wobei eine Hamming-Distanz von 2 zulässig ist.
Als weiteres Fehlermaß ist die Edit-Distanz implementiert, Abkürzung ed.

4 Ergebnisausgabe

Die Ergebnisausgabe besteht aus mehreren Teilen:

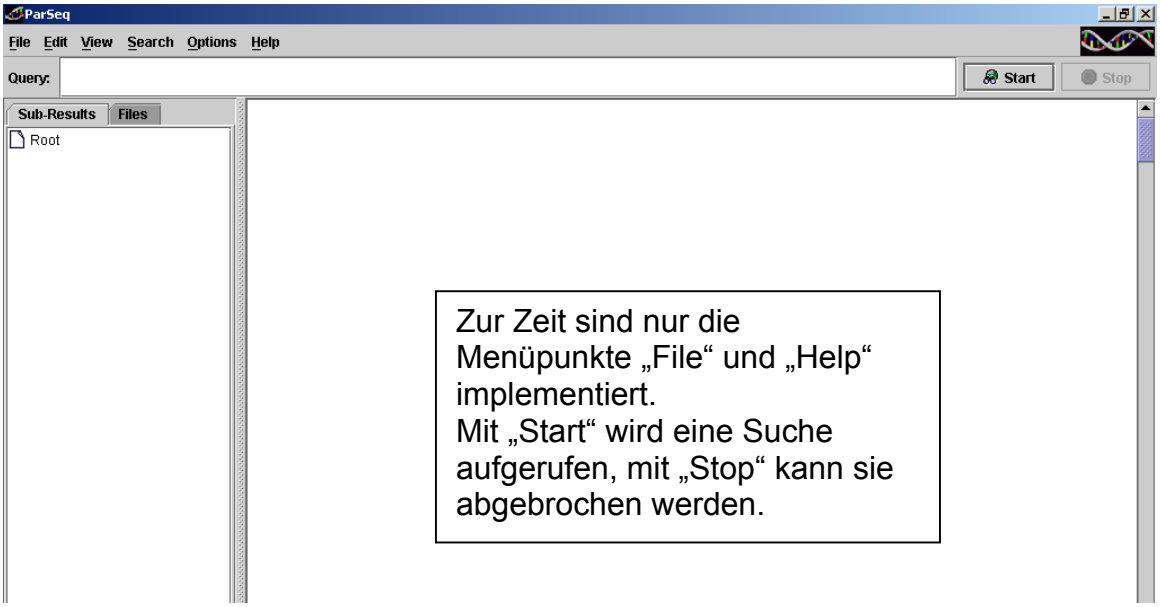
Zuerst erhält man Informationen über die durchgeführte Query:
Erst wird der eingegebene Suchausdruck angezeigt, dann der daraus berechnete reguläre Ausdruck (Regex). So wird bei einem zulässigen Fehler nach allen möglichen Varianten nacheinander gesucht. Diese werden aufgeführt.
Zusätzlich wird angegeben, in welcher Datei gesucht wurde.

Dann folgt die Liste der Treffer.
Angegeben ist jeweils die Treffersequenz, die Position auf der duchsuchten Zeichenfolge und bei Aminosäuresequenzen zusätzlich die Position des Treffers auf der DNA. Außerdem eine Trefferangabe mit Markierung der Struktur des Treffers. (Anfang und Ende der einzelnen Teilausdrücke)

Zuletzt folgt eine Statistik
Angegeben wird, wie viele Dateien welcher Größe durchsucht wurden, die Anzahl der Treffer und die Rechenzeit.

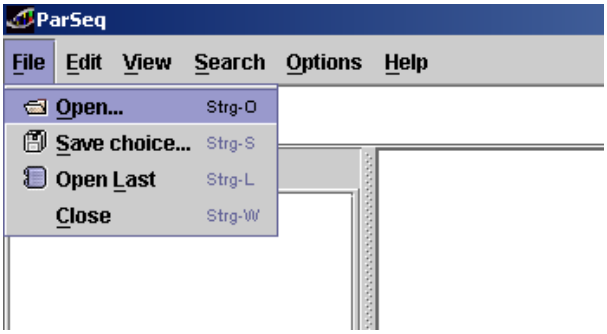
5 Benutzeroberfläche

Screenshot



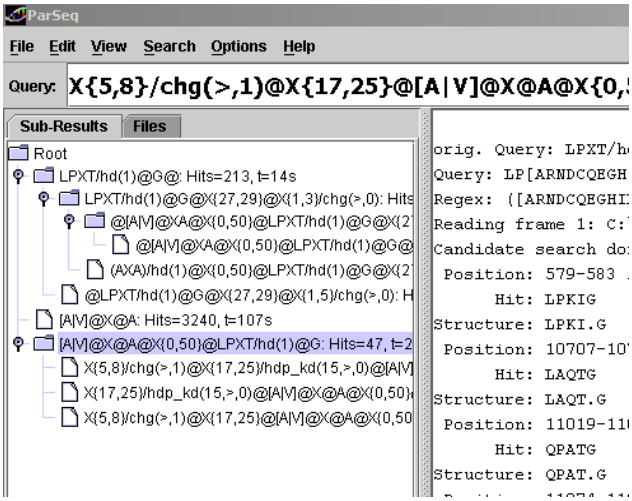
Zur Zeit sind nur die Menüpunkte „File“ und „Help“ implementiert.
Mit „Start“ wird eine Suche aufgerufen, mit „Stop“ kann sie abgebrochen werden.

Dateiauswahl



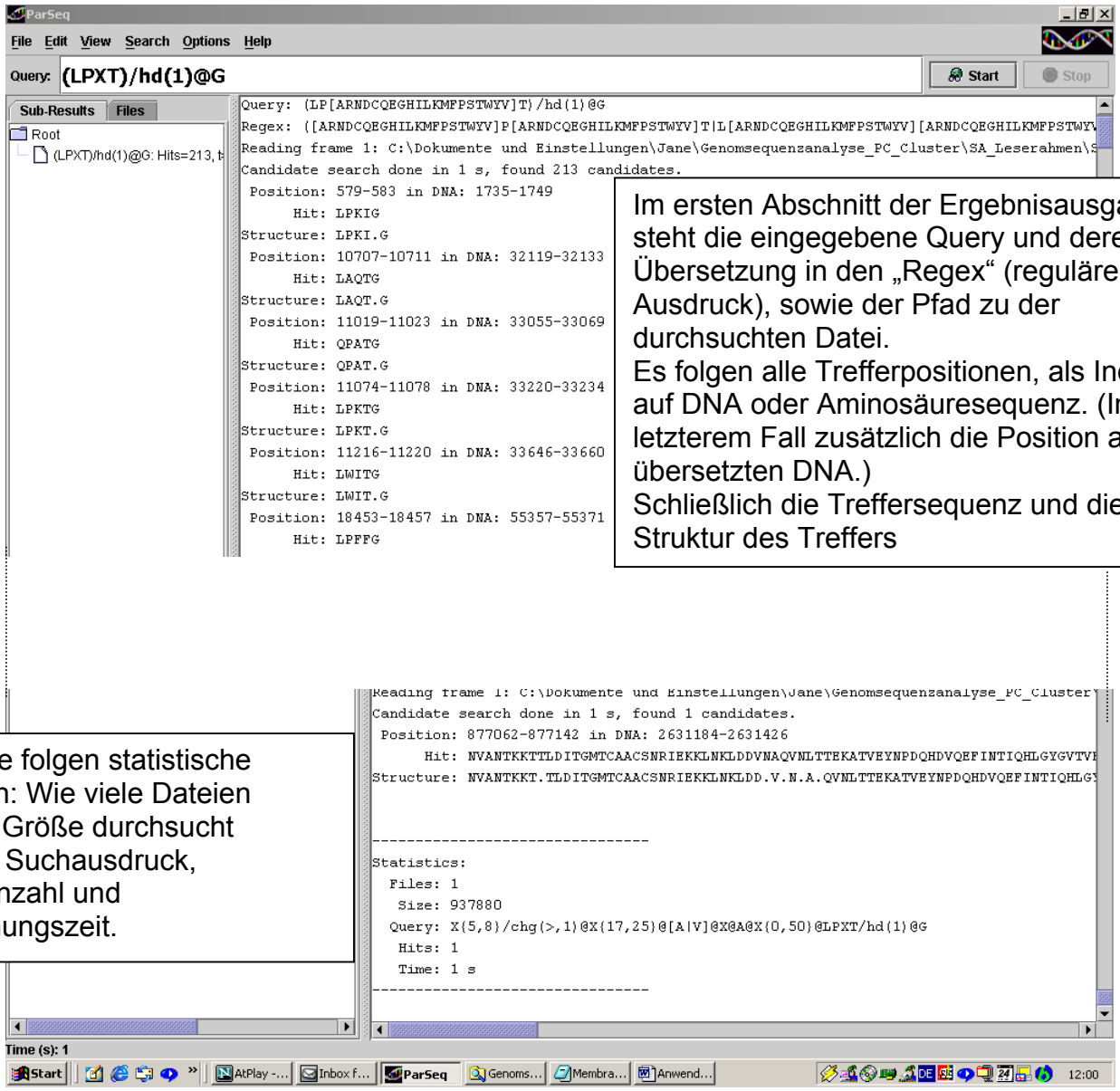
Es kann eine Textdatei aus dem Fileverzeichnis ausgewählt oder eine gespeicherte Auswahl wieder aufgerufen werden, oder die momentane Auswahl kann gespeichert werden. Abgefragt wird zudem, ob es sich um eine DNA- oder Aminosäuresequenz handelt und bei letzterer um welchen Leserahmen es sich handelt.

Eingabe einer Query



Als „Query“ wird der zu suchende reguläre Ausdruck eingetippt. Gesucht wird auf der ausgewählten Datei (root) oder man wählt per Anklicken aus dem Verzeichnisbaum das zu durchsuchende Zwischenergebnis aus.

Ergebnisausgabe



Im ersten Abschnitt der Ergebnisausgabe steht die eingegebene Query und deren Übersetzung in den „Regex“ (regulären Ausdruck), sowie der Pfad zu der durchsuchten Datei.
Es folgen alle Trefferpositionen, als Index auf DNA oder Aminosäuresequenz. (In letzterem Fall zusätzlich die Position auf der übersetzten DNA.)
Schließlich die Treffersequenz und die Struktur des Treffers

Am Ende folgen statistische Angaben: Wie viele Dateien welcher Größe durchsucht wurden, Suchausdruck, Trefferanzahl und Berechnungszeit.

6 Anhang

Liste der Nebenbedingungen

Hamming-Distanz: Symbol: **hd**
Argumente: Anzahl der Fehler

Beispiel:
LPXT/hd(1)@
Gesucht wird das Aminosäuremotiv LPXT, wobei eine Hammingdistanz von 1 zulässig ist.
(Hamming-Distanz von 1: Eine Abweichung an einer Position ist zulässig, also XPXT, oder LXXT, oder LPXX oder LPXT)

Edit-Distanz: Symbol: **ed**
Argumente: Anzahl der Fehler

Beispiel:
LPXT/ed(1)@
Gesucht wird das Aminosäuremotiv LPXT, wobei eine Editdistanz von 1 zulässig ist.
(Edit-Distanz von 1: Eine Abweichung an einer Position ist zulässig (wie bei der Hamming-Distanz) oder aber eine Deletion oder eine Insertion: zB: _PXT, LXPXT....)

Biochemische Eigenschaften: Symbole: siehe Tabelle im Anhang.
Argumente: (Klammerliste, durch Kommas getrennt):
Fenstergrösse (zwischen 1 und 25)
Vergleichsoperator (<,> oder=)
Notation: nach dem Teilmotiv, das die biochemische Eigenschaft aufweisen soll, einen Schrägstrich setzen und danach das entsprechende Symbol und die Argumentliste einfügen.

Der Wertebereich richtet sich nach der zu verwendenden Tabelle. Die Werte der einzelnen Aminosäuren können für alle implementierten Scalingtabellen unter folgender Adresse im Internet nachgeschlagen werden:

http://us.expasy.org/cgi-bin/protscale.pl

Beispiel:
X{5,8}/chg(>,2)@X{19,23}/hdp_kd(5,>,-1)@
Gesucht wird ein Abschnitt von 5-8 frei wählbaren Aminosäuren, mit einer positiven Ladung größer 2 (Summe) und ein sich anschließender Bereich von 19 bis 23 frei wählbaren Aminosäuren, die innerhalb der Fenstergrösse von 5 im Mittel einen Hydrophobizitätsscore nach Kyte-Doolittle von größer –1 haben.

Verwendete Scaling-Listen und Symbole

Bezeichnung	Autor	Artikel	Symbol
Hydropathicity	Kyte J., Doolittle R.F	J. Mol. Biol. 157:105-132(1982)	**hdp_kd**
Mokecular Weight		Most textbooks	**mcw_**
Bulkiness	Zimmerman J.M., Eliezer N., Simha R.	J. Theor. Biol. 21:170-201(1968)	**blk_ze**
Polarity	Grantham R	Science 185 :862-864(1974)	**pol_g**
Recognition factors	Fraga S.	Can. J. Chem. 60:2606-2610(1982)	**ref_f**
Optimized matching hydrophobicity (OMH)	Sweet R.M., Eisenberg D	J. Mol. Biol. 171:479-488(1983)	**omh_se**
Hydrophobicity (delta G1/2 cal)	Abraham D.J., Leo A.J	Proteins: Structure, Function and Genetics 2:130-152(1987)	**hdp_al**
Hydrophobicity (free energy of transfer to surface in kcal/mole)	Bull H.B., Breese K	Arch. Biochem. Biophys. 161:665-670(1974)	**hdp_bb**
Hydrophobicity scale based on free energy of transfer (kcal/mole)	Guy H.R.	Biophys J. 47:61-70(1985)	**hdp_g**
Hydrophobicity scale (contact energy derived from 3D data)	Miyazawa S., Jernigen R.L	Macromolecules 18:534-552(1985)	**hdp_mj**
Hydrophobicity scale (pi-r)	Roseman M.A	J. Mol. Biol. 200:513-522(1988)	**hdp_r**
Antigenicity value X 10	Welling G.W., Weijer W.J., Van der Zee R., Welling-Wester S	FEBS Lett. 188:215-218(1985)	**agv_wv**

Bezeichnung	Autor	Artikel	Symbol
Hydrophilicity scale derived from HPLC peptide retention times	Parker J.M.R., Guo D., Hodges R.S	Biochemistry 25:5425-5431(1986)	**hdp_pg**
Hydrophobicity indices at ph 7.5 determined by HPLC	Cowan R., Whittaker R.G	Peptide Research 3:75-80(1990)	**hdp_cw**
Retention coefficient in HFBA	Browne C.A., Bennett H.P.J., Solomon S	Anal. Biochem. 124:201-208(1982)	**rec_bb**
Retention coefficient in HPLC, pH 2.1.	Meek J.L	Proc. Natl. Acad. Sci. USA 77:1632-1636(1980)	**rec_m**
Molar fraction (%) of 2001 buried residues	Janin J.		**mfr_j**
Proportion of residues 95% buried (in 12 proteins)	Chothia C.	J. Mol. Biol. 105:1-14(1976).	**prb_c**
Atomic weight ratio of hetero elements in end group to C in side chain	Grantham R.	Science 185 :862-864(1974)	**whe_g**
Average flexibility index	Bhaskaran R., Ponnuswamy P.K.	Int. J. Pept. Protein. Res. 32 :242-255(1988)	**afi_bp**
Conformational parameter for beta-sheet	Chou P.Y., Fasman G.D	Adv. Enzym. 47:45-148(1978)	**pbs_cf**

Bezeichnung	Autor	Artikel	Symbol
Charge		http://speedy.embl-heidelberg.de/aas/	**chg_**
Conformational parameter for alpha helix	Deleage G., Roux B.	Protein Engineering 1:289-294(1987)	**pah_dr**
Conformational parameter for beta-turn	Deleage G., Roux B.	Protein Engineering 1:289-294(1987)	**pbt_dr**
Normalized frequency for alpha helix	Levitt M	Biochemistry 17:4277-4285(1978)	**fah_l**
Normalized frequency for beta-turn	Levitt M.	Biochemistry 17:4277-4285(1978)	**fbt_l**
Conformational preference for antiparallel beta strand	Lifson S., Sander C.	Nature 282:109-111(1979).	**abz_ls**
Overall amino acid composition (%).	McCaldon P., Argos P.	Proteins: Structure, Function and Genetics 4:99-122(1988)	**aac_ca**
Relative mutability of amino acids (Ala=100)	Dayhoff M.O., Schwartz R.M., Orcutt B.C	“Atlas of Protein Sequence and Structure”, Vol.5, Suppl.3 (1978)	**maa_ds**
Polarity	Zimmerman J.M., Eliezer N., Simha R.	J. Theor. Biol. 21:170-201(1968)	**pol_ze**
Refractivity	Jones. D.D.	J. Theor. Biol. 50:167-184(1975)	**ref_j**
Conformational parameter for alpha helix (computed from 29 proteins).	Chou P.Y., Fasman G.D.	Adv. Enzym. 47:45-148(1978)	**pah_cf**
Conformational parameter for beta-turn(computed from 29 proteins).	Chou P.Y., Fasman G.D.	Adv. Enzym. 47:45-148(1978)	**pbt_cf**

Bezeichnung	Autor	Artikel	Symbol
Conformational parameter for beta-sheet.	Deleage G., Roux B.	Protein Engineering 1:289-294(1987)	**pbs_dr**
Conformational parameter for coil	Deleage G., Roux B.	Protein Engineering 1:289-294(1987).	**pco_dr**
Normalized frequency for beta-sheet	Levitt M	Biochemistry 17:4277-4285(1978)	**fbs_l**
Conformational preference for total beta strand (antiparallel+parallel)	Lifson S., Sander C.	Nature 282:109-111(1979)	**pbz_ls**
Membrane buried helix parameter	Rao M.J.K., Argos P	Biochim. Biophys. Acta 869:197-214(1986)	**mbh_ra**

Einbuchstabencode der Aminosäuren

Name	Symbol, entspricht 1-Buchstabencode	3-Buchstabencode
Glycin	G	Gly
Alanin	A	Ala
Valin	V	Val
Leucin	L	Leu
Isoleucin	I	Ile
Cystein	C	Cys
Methionin	M	Met
Phenylalanin	F	Phe
Tyrosin	Y	Tyr
Tryptophan	W	Trp
Prolin	P	Pro
Serin	S	Ser
Threonin	T	Thr
Asparagin	N	Asn
Glutamin	Q	Gln
Asparaginsäure	D	Asp
Glutaminsäure	E	Glu
Histidin	H	His
Lysin	K	Lys
Arginin	R	Arg

Abkürzungen zur Beschreibung von Zeichenkombinationen in regulären Ausdrücken

Aminosäurekombinationen

X=[ARNDCQEGHILKMFPSTWYV]

Nukleotidkombinationen

X=[AGTC]
N=[AGCT]
R=[GA]
Y=[TC]
K=[GT]
M=[AC]
S=[GC]
W=[AT]
H=[ACT]
B=[GTC]
V=[GCA]
D=[GAT]

Geplante Erweiterungen

Boolesche Operatoren:
Teilausdrücke sollen durch Boole'sche Operatoren verbunden werden können. Eingeführt werden UND, NICHT und ODER.
Dadurch soll ermöglicht werden, mehrere Bedingungen an einen Teilausdruck zu stellen, sowie einzugeben, welche Zeichen oder Zeichenfolgen in einem Teilausdruck gar nicht vorkommen dürfen.

Fehler:
Fehler sollen nicht nur für Teilausdrücke zulässig sein, sondern auch für strukturierte Motive, beispielsweise werde nach MOTIV1-variabler Abstand-MOTIV2 gesucht, ein Fehler soll zulässig sein, entweder in Motiv1 oder in Motiv2.
Weiter soll es möglich sein, Fehler innerhalb einer Sequenz mit bestimmten biologischen Eigenschaften zuzulassen. So etwa eine Abweichung vom geforderten Grenzwert in einer definierten Anzahl von Fensterpositionen.

Statistische Maße:
Bislang werden zur Berechnung der Grenzwerte innerhalb einer Fensterposition ausschließlich die Mittelwerte berechnet. Damit Einzelwerte kein zu hohes Gewicht erhalten, sollten auch andere statistische Maße (Standardabweichung...) verwendet werden können.

Translation:
Das Programm soll in der Lage sein, eine DNA-Sequenz in die sechs möglichen Aminosäuresequenzen zu übersetzen (sechs Leserahmen).

Einbau von ORF-Grenzen:
Zusätzlich soll eine Liste von ORF-Grenzen dem Programm mitgegeben werden. Im regulären Ausdruck kann dann ergänzend angegeben werden, wie weit das zu suchende Motiv von Beginn oder Ende der Sequenz entfernt sein darf.
Biologischer Hintergrund: Manche funktionstragenden Teile von Proteinen, beispielsweise Signalsequenzen, liegen in definierter Nähe zum C- oder N-terminalen Ende. Regulatorische Elemente auf der DNA liegen in bestimmtem Abstand zu den ORF-Grenzen. Diese Informationen sollen in die Suche mit eingehen können.

Komplementäre DNA:
Integriert werden soll weiter eine automatische Suche von Motiven auch auf dem komplementären DNA-Strang und die Möglichkeit, Treffer auf dem kodierenden DNA-Strang zu denen auf dem komplementären DNA-Strang in räumlichen Bezug zu setzen.
Biologischer Hintergrund: Bindungsstellen für Proteine auf der DNA können dadurch gekennzeichnet sein, dass mehrere Sequenzmotive auf kodierendem DNA-Strang und komplementärem Strang in direkter Nachbarschaft zueinander liegen.

Verzeichnisse durchsuchen:
Es soll ermöglicht werden, viele DNA-Sequenzen nacheinander zu durchsuchen. Eingabe könnte beispielsweise die Adresse eines Ordners sein, alle darin befindlichen Textfiles sollen dann nach dem eingegebenen regulären Ausdruck durchsucht werden.
Biologischer Hintergrund: Es sollen auch Verzeichnisse eukaryotischer Gene durchsucht werden können.

Sortierte Trefferliste (Ranking):
Gewichtung der gefundenen Treffer nach „Maß, in dem die Anforderungen erfüllt werden“ oder „Kommen erlaubte Fehler vor, oder liegt das Motiv in idealer Form vor?“
Die gefundenen Treffer sollen gemäß dieser Gewichtung sortiert ausgegeben werden.

Integration der ParSeq-Treffer in eine annotierte Darstellung des untersuchten Genoms
Ein Viewer zur Darstellung von Genbank®[1]-Einträgen wird geschrieben. In diese Darstellung sollen ParSeq-Treffer interaktiv integriert werden können. Dadurch soll ermöglicht werden, Annotationsdaten und Sequenzen zu gefundenen ORFs durch anklicken abfragen zu können.

Verwendung weiterer Maße zur Untersuchung der biochemischen Eigenschaften
Verwendung von „weighted matrices“ statt Scalingtabellen.
Überprüfung durch spezialisierte Programme.

Einbindung externen Programme
Weitergabe von Treffern an externe Programme, die in der Lage sind, die Treffer auf bestimmte Eigenschaften hin zu überprüfen, etwa durch Verwendung von neuronalen Netzen.
Biologischer Hintergrund: Die Einschätzung von biochemischen Eigenschaften allein Aufgrund der Aminosäuresequenz ist nur in wenigen Fällen ausreichend. Für viele Fragestellungen stehen bereits spezialisierte Programme zur Verfügung, die durch komplexere Ansätze bessere Ergebnisse liefern können. Diese Programme sind aber meist nicht in der Lage, lange Sequenzen zu durchsuchen. Mit ihnen können aber die Treffersequenzen bearbeitet werden.

Beschleunigung der Rechenzeit
Parallelisierung der Algorithmen und Implementierung auf dem Kepler-Cluster.[2]

[1] Oder Annotationsdaten aus anderen Datenbanken im GenBank® -Format.
(GenBank®: Gendatenbank des NCBI – National Center for Biotechnology Information, USA)
[2] http://kepler.sfb382-zdv.uni-tuebingen.de/kepler/start.shtml