Matthias Jauernig
06INM

Leipzig, 20.12.06

LV Programmverifikation
Prof. Dr. Uwe Petermann

Vorstellung der Projektaufgabe „Genotyp – Phänotyp“

Bei Gene Expression Programming (GEP[1]) sollen Computerprogramme unter Zuhilfenahme evolutionärer Algorithmen entwickelt werden. Die Individuen bestehen hier aus Genotypen und Phänotypen. Genotypen stellen in der Biologie das Erbgut eines Organismus dar, während durch den Phänotyp dessen Erscheinungsbild festgelegt wird.

Ähnlich verhält es sich bei GEP. Die Genotypen sind Chromosomen fixer Länge, welche aus einem oder mehreren Genen gleicher Länge bestehen können. Die Phänotypen stellen die Repräsentation der Chromosomen als Ausdrucksbäume (expression tree, ET) dar. Genotypen fixer Länge können Phänotypen unterschiedlicher Länge erzeugen.

Kurz zum Aufbau eines Chromosoms im Sinne des GEP: es besitzt eine feste Länge und besteht aus einzelnen Zeichen, die entweder n-stellige Funktionssymbole (FS) oder Terminalsymbole (TS) darstellen. Ein Chromosom als Ganzes ist somit eine Zeichenkette, kann logisch jedoch in mehrere Gene gleicher fester Länge unterteilt werden. Solch ein Gen ist somit wiederum eine Zeichenkette aus FS und TS und besteht strukturell aus einem Kopf sowie einem Rest. Im Kopf können FS und TS beliebig aneinander gereiht vorkommen, wohingegen im Rest nur TS erlaubt sind. Zwischen der Länge des Kopfes h und der Länge des Restes t besteht folgende Beziehung (n sei dabei die höchste Arität eines FS):

$$t = h*(n-1) + 1$$

Dadurch wird gesichert, dass jedes Gen mit einer solchen Struktur durch einen ET dargestellt werden kann. Die Länge des Kopfes ist a priori festzulegen, wodurch die Länge des Restes und schließlich auch die Länge des ganzen Genes berechnet werden kann.

Das folgende Beispiel illustriert ein Chromosom mit 3 Genen, deren Kopf jeweils die Länge 4 haben. Die Menge der Funktionssymbole sei gegeben durch FS={/,*,+,Q}, wobei die arithmetischen Operationen und mit Q die Bildung der Quadratwurzel gemeint sind. Die höchste Arität ist hier n=2. Die Menge der Terminalsymbole bestehe aus TS={a,b}. Aus o.a. Formel ergibt sich für den Rest eine Länge von t=5 und somit eine Genlänge von t+h=9 sowie eine Chromosomlänge von 3*9=27:

Chromosom																									
Gen 1									Gen 2								Gen 3								
Kopf1				Rest1					Kopf2				Rest2					Kopf3				Rest3			
/	*	a	Q	***a***	***b***	***a***	***a***	***a***	+	a	b	/	***a***	***a***	***b***	***b***	***b***	*	Q	+	/	***b***	***b***	***a***	***b***	***a***

Der Phänotyp lässt sich aus dieser Darstellung als ET gewinnen. Speziell kann für jedes Gen eines Chromosoms ein eigener Unterbaum (Sub-ET) erzeugt werden (solche Unterbäume

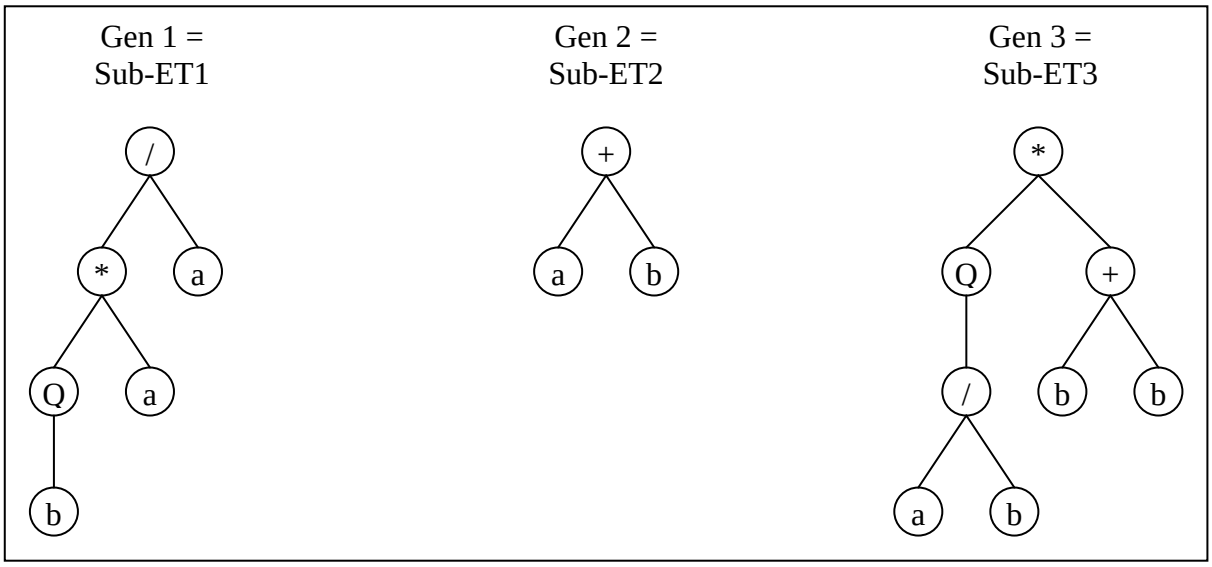
[1] Siehe: Candida Ferreira, http://www.gene-expression-programming.com

lassen sich dann wieder zusammenfügen, doch dies soll hier nicht weiter Bestandteil der Betrachtung sein).

Ein Unterbaum lässt sich durch einen systematischen Aufbau aus einem Gen erstellen, indem jedes Symbol eines Gens von links startend betrachtet wird. Das erste Symbol des Gen-Kopfes stellt die Baumwurzel dar. Hat es die Arität n, so müssen sich n Kinder anschließen, die nächsten n Symbole des Gens sind dafür der Reihe nach zu verwenden. Dies hat rekursiv zu erfolgen, sodass sich ein Baum von oben nach unten und in jeder Ebene von links nach rechts aufbaut. Die Prozedur wird beendet, wenn sich nur noch TS in den Baumblättern befinden. Da TS die Arität 0 besitzen und im Rest eines Gens nur TS vorhanden sind, terminiert der Algorithmus mit einem Baum, der maximal so viele Knoten hat wie das Gen Zeichen besitzt. Minimal ergibt sich ein Baum mit nur einem Knoten in dem Fall, wenn im ersten Zeichen des Gen-Kopfes ein TS steht.

Es müssen beim Aufbau eines Sub-ET also nicht alle Zeichen des zugehörigen Gens benutzt werden. Die restlichen Zeichen sind dann zu dem betrachteten Zeitpunkt überflüssig, können durch Evolution des Gen-Kopfes aber in späteren Generationen Verwendung finden. Durch die Abbruchbedingung „alle Blätter sind TS“ sind somit für ein Gen fester Länge variable Baumgrößen vorstellbar, was diesen Algorithmus flexibel einsetzbar macht.

Für das oben angeführte Beispiel ergeben sich die folgenden Sub-ETs:



Die Aufgabe beinhaltet die Spezifikation der Begriffe Genotyp und Phänotyp als Datenstrukturen im KIV-System. Weiterhin soll die Erstellung von ETs aus Chromosomen heraus in KIV formalisiert werden (und möglicherweise auch die Gegenrichtung). Dies erfordert folgende Schritte bzw. Überlegungen:

- Ein Genotyp soll zunächst nur aus einem Gen bestehen. Eine Erweiterung auf beliebig viele Gene könnte sich im Nachhinein anschließen.
- Zur Spezifikation von „Genotyp“: dieser sollte sich darstellen lassen als Liste von Symbolen. Jeder Symbol besitzt ein Zeichen und eine Arität. Terminalsymbole sollen als 0-stellige Funktionssymbole dargestellt werden.
- Zur Spezifikation von „Phänotyp“: dieser sollte sich darstellen lassen als beliebiger n-ärer Baum. Es ist zu prüfen ob die KIV-Bibliothek eine solche Datenstruktur bereitstellt.
- Die Umwandlung Genotyp <=> Phänotyp ist rekursiv zu definieren und zu implementieren.