

Scalaの文字列処理

Day 3 コードポイントとサロゲートペア

コードポイント

文字単位を正確に扱いたい場合は、Charではなくコードポイントを使用する。コードポイントは、Unicode上での番地を意味し、この符号化方式はUTF-32と呼ばれる。

プログラム上で文字を扱う場合は、Byte Order Markはつけず、ビッグエンディアンで扱う。

	符号化方式	実装	容量
Java/Scalaの Code Point	UTF-32BE	Int	4,294,967,296(32bits)
Java/Scalaの char/Char	UTF-16BE	BMP領域の文字 = Char 1 つ	65,536(16bits)
		追加領域の文字 = Char 2 つ	4,294,967,296(32bits)
C/C++のchar	Latin1	char	256(8bits)
	UTF-32BE	Windows上でのwchar_t	4,294,967,296(32bits)
	UTF-16BE	Unix上でのwchar_t	65,536(16bits)
	UTF-16BE	char16_t	65,536(16bits)
	UTF-32BE	char32_t	4,294,967,296(32bits)

サロゲートペア

追加領域にある 1 文字を 2 文字で表現する機構

これらの 2 文字の組をサロゲートペアと呼び、構成する文字の前方を上位サロゲートと後方を下位サロゲートと呼ぶ

	領域	容量
追加領域	[U+10000, U+10FFFF]	1,048,576 (20 bits)
上位サロゲート	[U+D800, U+DBFF]	1,024 (10 bits)
下位サロゲート	[U+DC00, U+DFFF]	1,024 (10 bits)

サロゲートペア

追加領域にある 1 文字を 2 文字で表現する機構

これらの 2 文字の組をサロゲートペアと呼ぶ
上位サロゲートと後方を下位サロゲート

10000111111111111111(2)
左から1が1個, 0が4個, 1が16個
合計21個=21bits

	領域	容量
追加領域	[U+10000, U+10FFFF]	1,048,576 (20 bits)
上位サロゲート	[U+D800, U+DBFF]	1,024 (10 bits)
下位サロゲート	[U+DC00, U+DFFF]	1,024 (10 bits)

サロゲートペア

追加領域にある 1 文字を 2 文字で表現する機構

これらの 2 文字の組をサロゲートペアと呼ぶ
上位サロゲートと後方を下位サロゲート

10000111111111111111(2)
左から1が1個, 0が4個, 1が16個
合計21個=21bits

	領域	容量
追加領域	[U+10000, U+10FFFF]	1,048,576 (20 bits)
上位サロゲート	[U+D800, U+DBFF]	1,024 (10 bits)
下位サロゲート	[U+DC00, U+DFFF]	1,024 (10 bits)

16bitsのCharでは追加領域の文字(21bits)をChar 1 つで表現不可能

サロゲートペア

追加領域にある 1 文字を 2 文字で表現する機構

これらの 2 文字の組をサロゲートペアと呼ぶ
上位サロゲートと後方を下位サロゲート

10000111111111111111(2)
左から1が1個, 0が4個, 1が16個
合計21個=21bits

	領域	容量
追加領域	[U+10000, U+10FFFF]	1,048,576 (20 bits)
上位サロゲート	[U+D800, U+DBFF]	1,024 (10 bits)
下位サロゲート	[U+DC00, U+DFFF]	1,024 (10 bits)

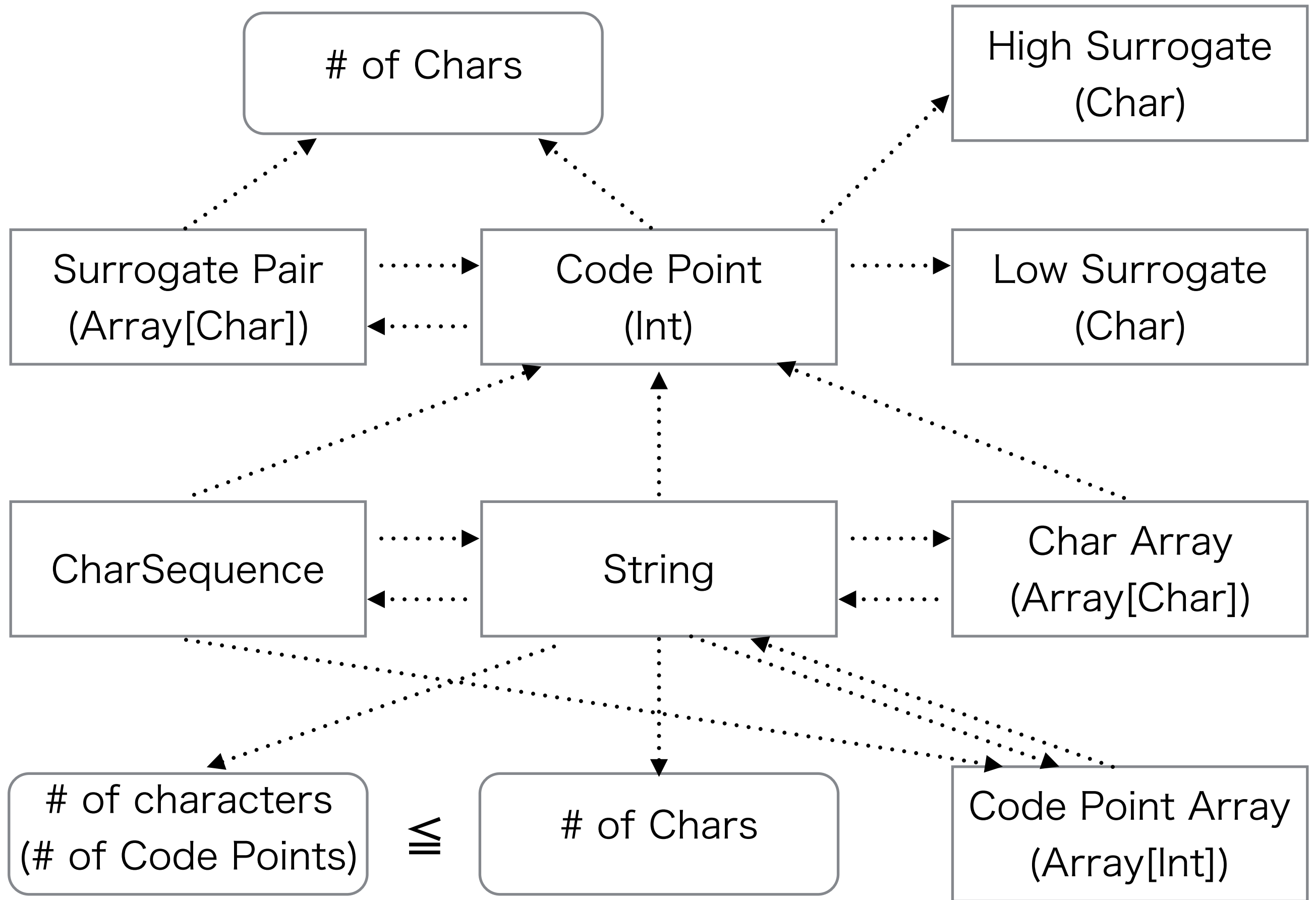
16bitsのCharでは追加領域の文字(21bits)をChar 1 つで表現不可能
→サロゲートペアに変換しChar 2 つの32bitsで扱う

コードポイントと サロゲートペアの変換方法

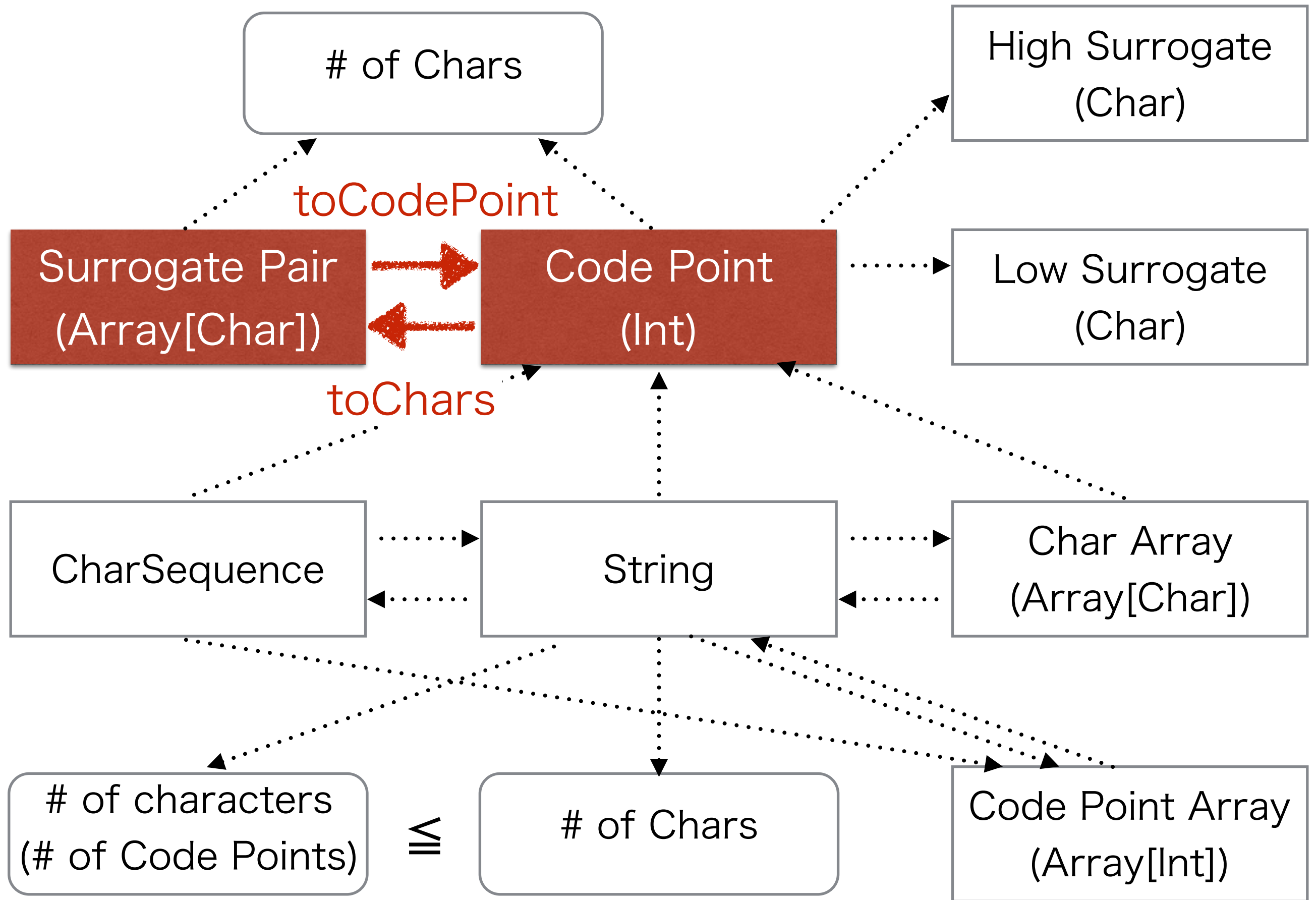
コードポイント = $0x10000 + (\text{上位サロゲート} - 0xD800) * 0x400 + (\text{下位サロゲート} - 0xDC00)$

上位サロゲート = $(\text{コードポイント} - 0x10000) / 0x400 + 0xD800$

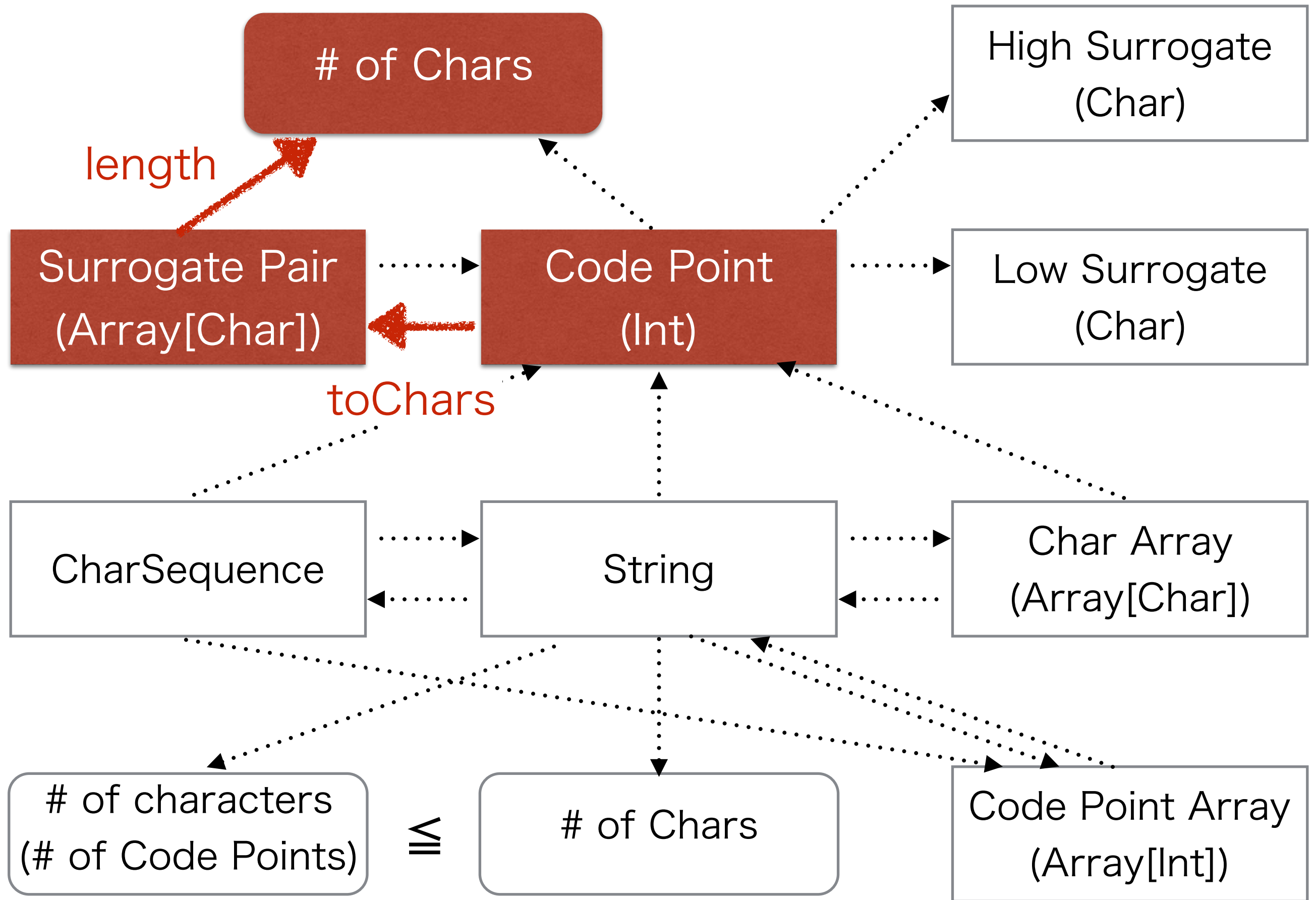
下位サロゲート = $(\text{コードポイント} - 0x10000) \% 0x400 + 0xDC00$



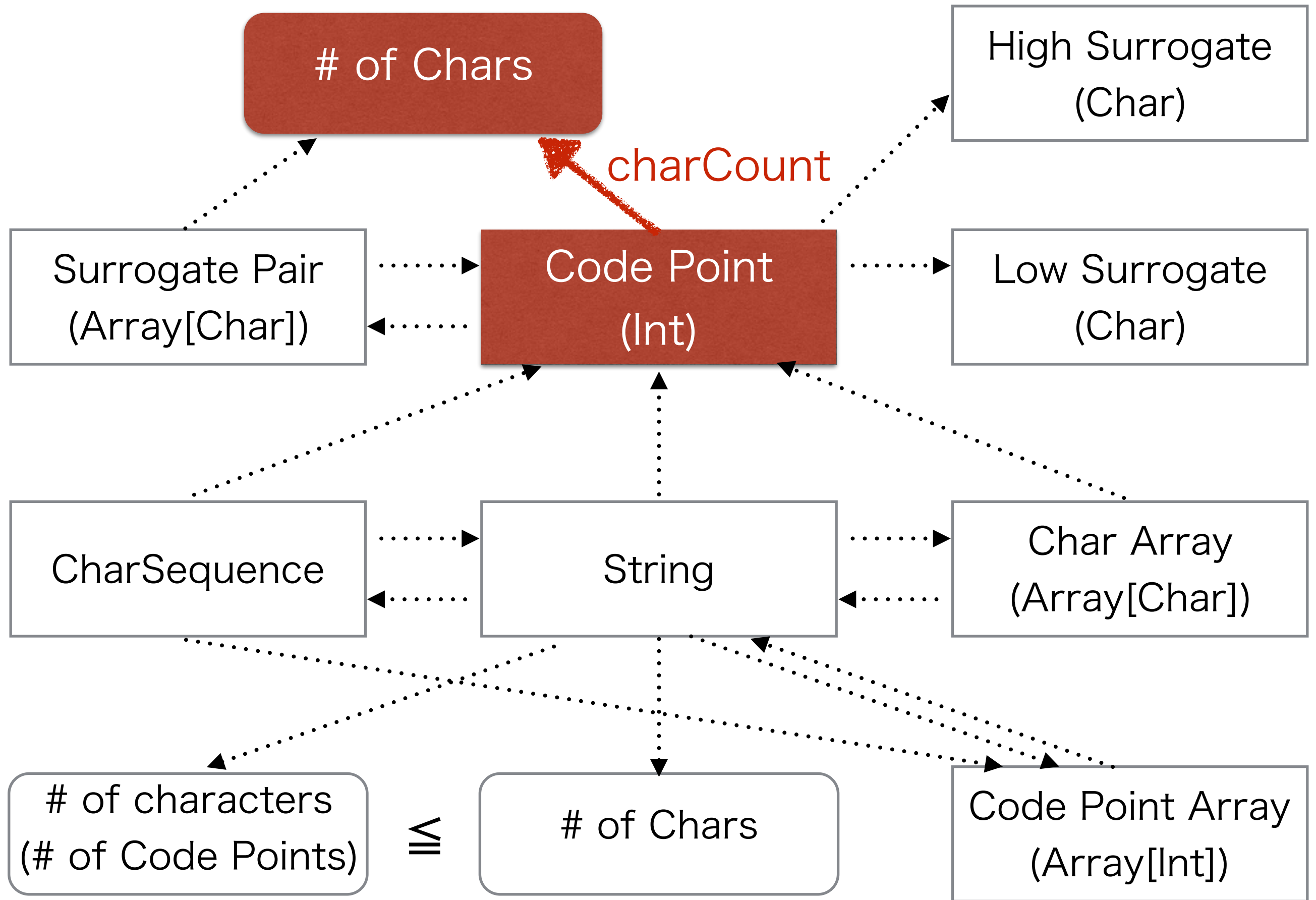
Surrogate Pair が含まれている場合は一致せず。



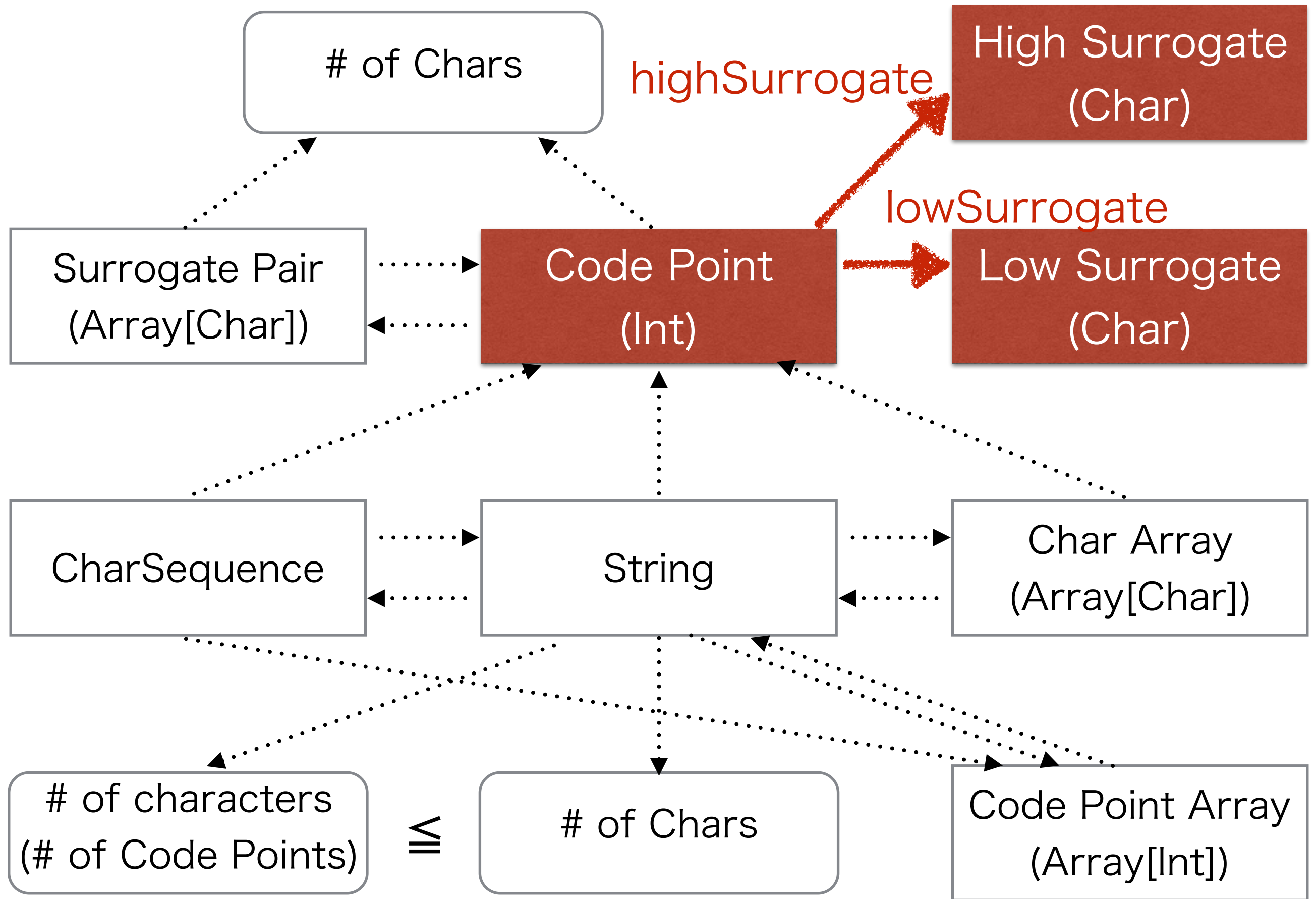
Surrogate Pair が含まれている場合は一致せず。



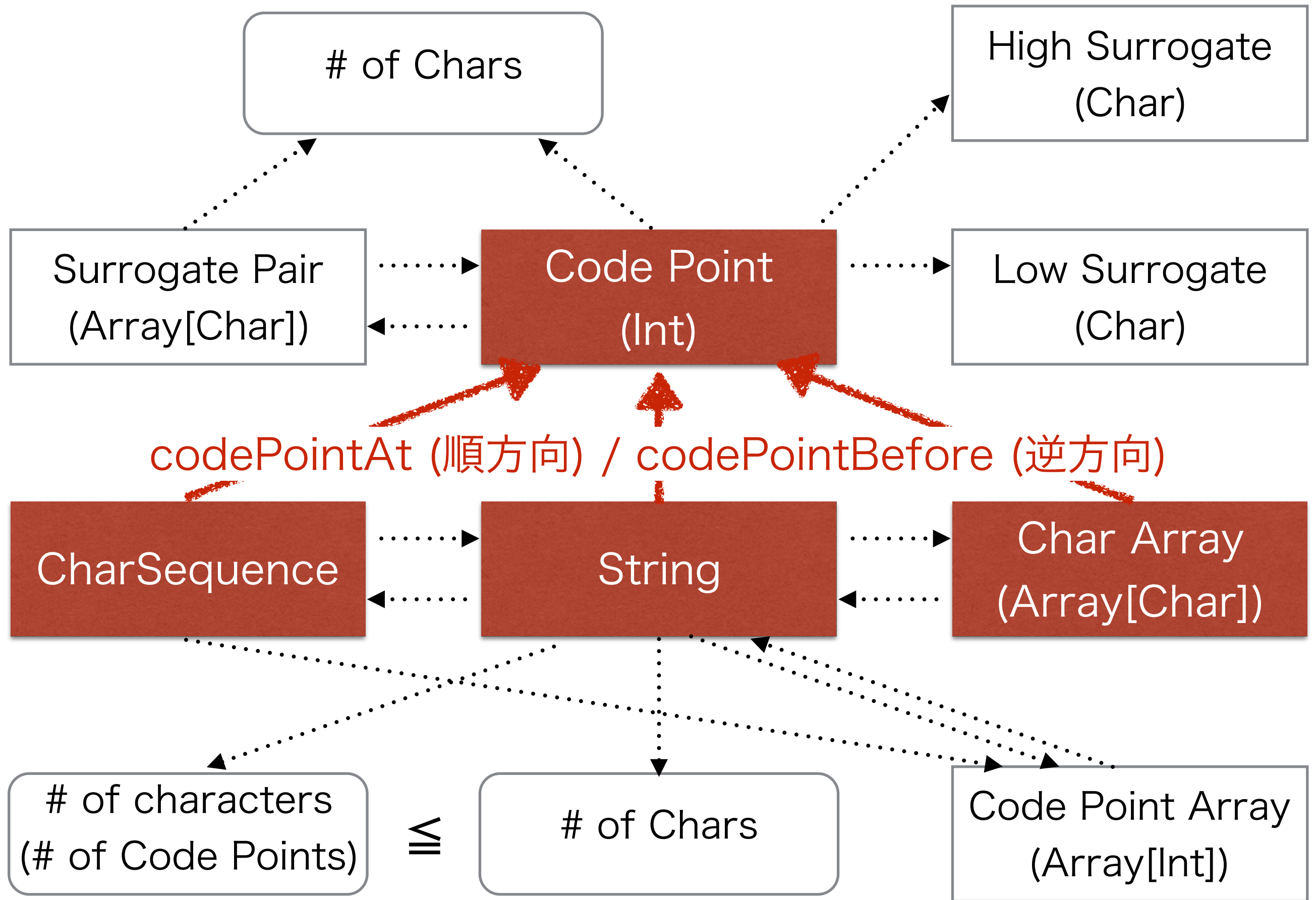
Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。

指定インデックスにある文字のコードポイントの取得方法

- ・ 指定インデックスにある文字のコードポイントを取得（順方向に解析）
- ・ 指定インデックスの**一つ前**にある文字のコードポイントを取得（逆方向に解析）

方向\入力	Char配列	CharSequence	String
順方向(前方から後方)	Character.codePointAt		str.codePointAt
逆方向(後方から前方)	Character.codePointBefore		str.codePointBefore

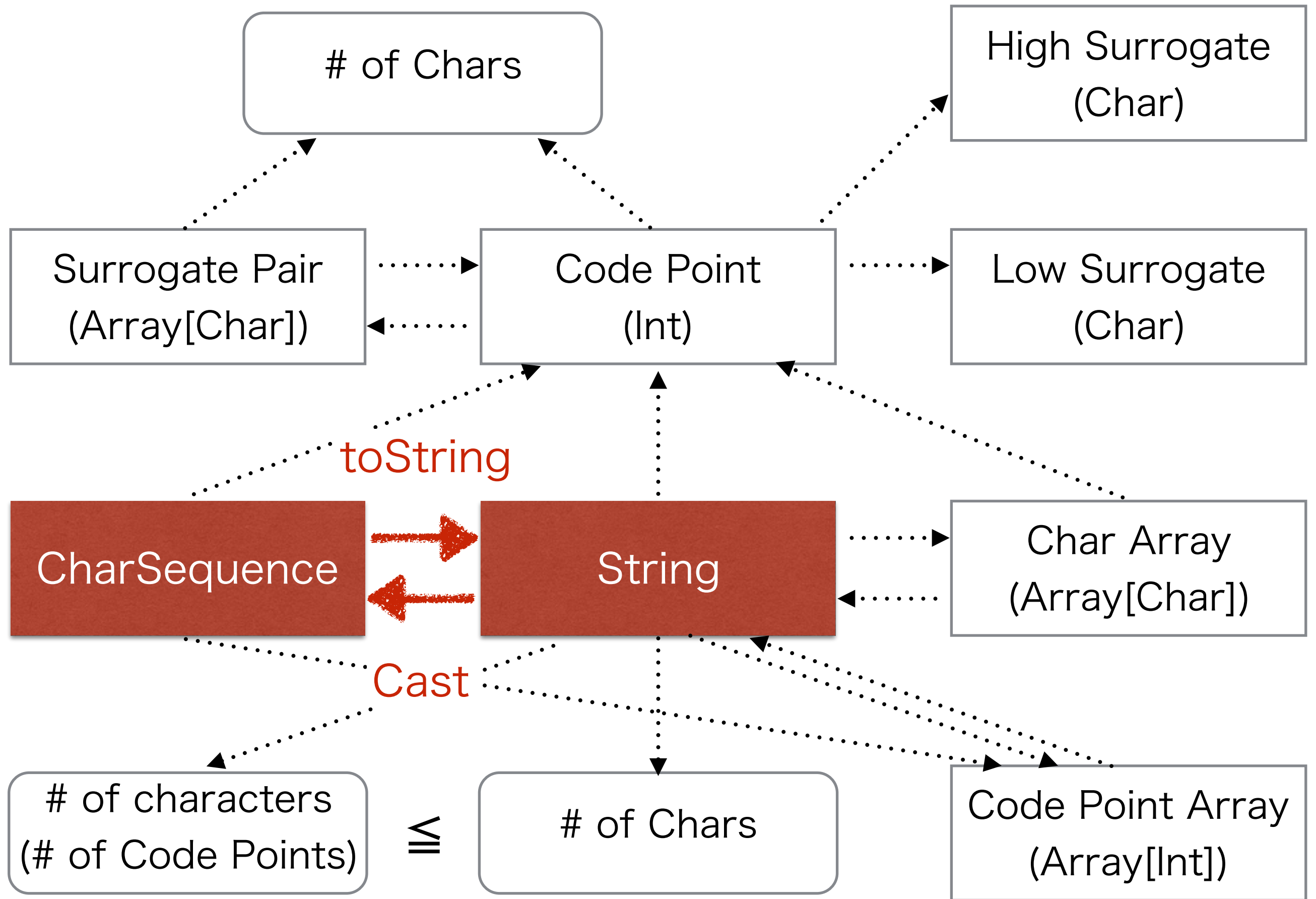
CharSequenceやStringでは、それらが内部に持つChar配列にアクセスして処理

指定インデックスにある文字を1つだけコードポイントで取り出す場合は上記で問題ないが、CharSequenceやStringの持つCharの**すべて**をコードポイントに変換したい場合、CharSequenceやStringを介して内部のChar配列にアクセスするのは遠回りで処理が低速

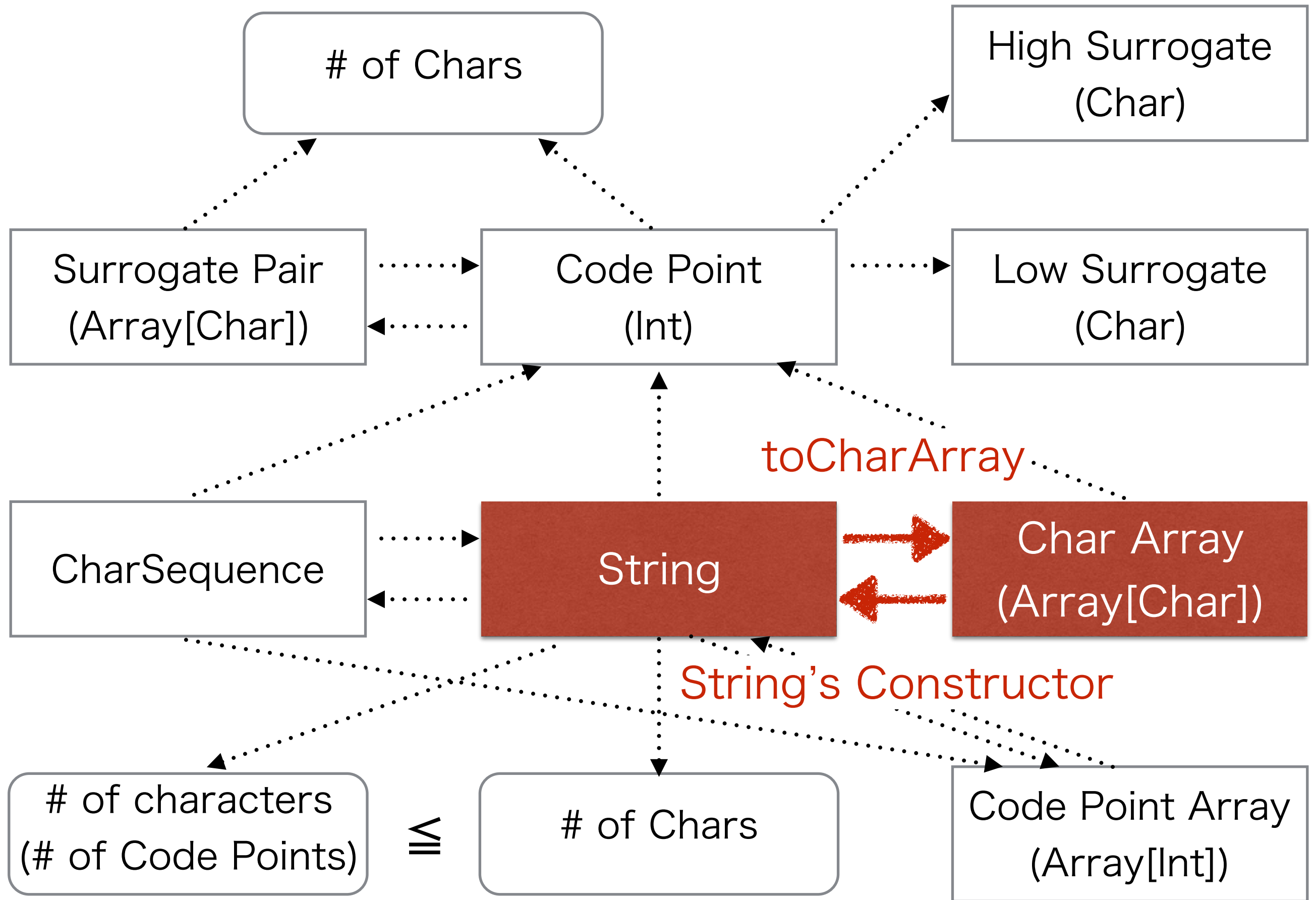
サロゲートペアに対する codePointAt/codePointBeforeの挙動

input	Character. codePointAt(input, 0)	Character. codePointBefore(input, input.length)
Array[Char] (0xD842, 0xDFB7)	0x20BB7	0x20BB7
Array[Char] (0xD842)	0xD842	0xD842
Array[Char] (0xDFB7)	0xDFB7	0xDFB7

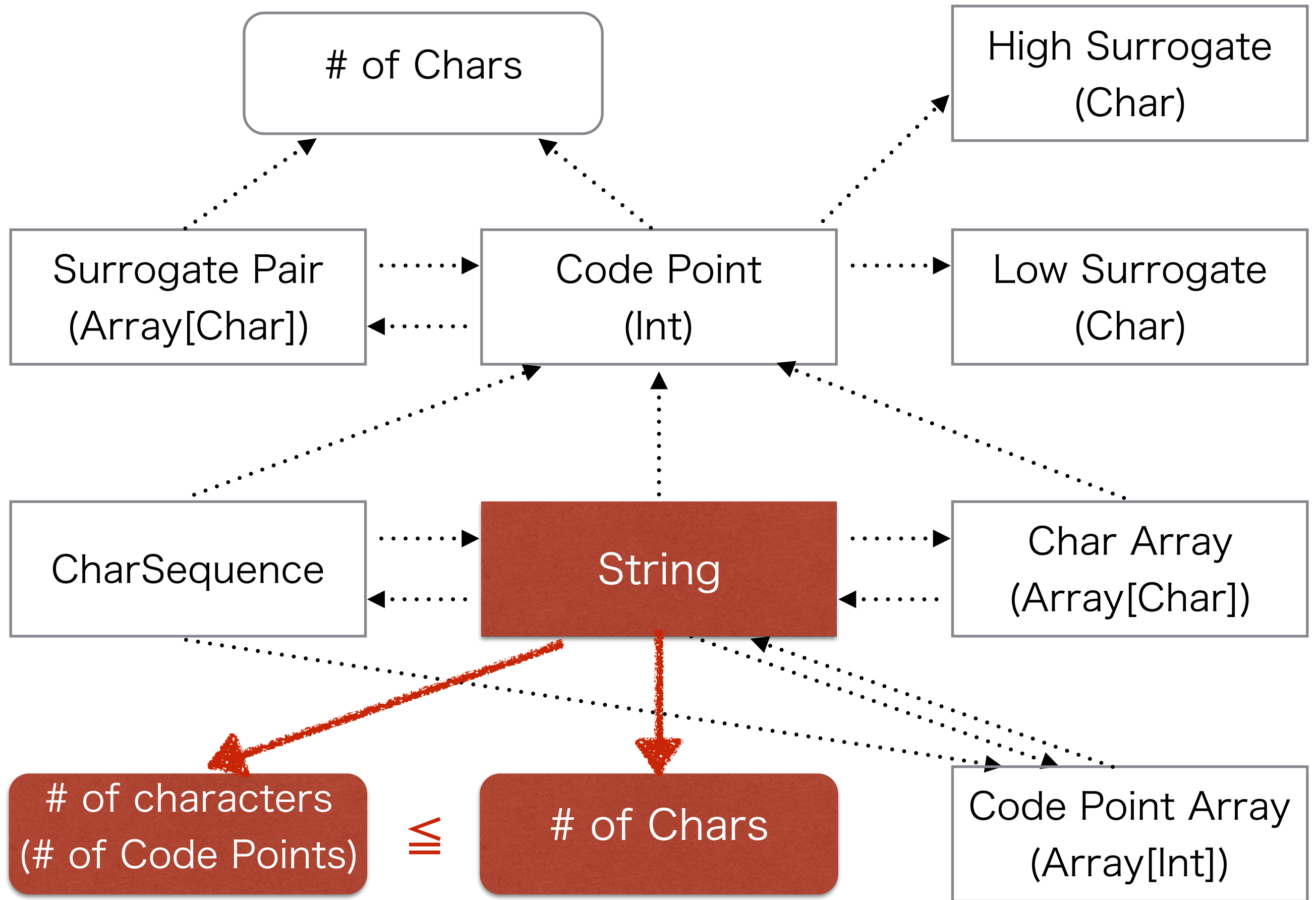
val input = “吉”//0xD842, 0xDFB7	出力
input.codePointAt(0)	0x20BB7
input.codePointAt(1)	0xDFB7
input.codePointBefore(2)	0x20BB7
input.codePointBefore(1)	0xD842



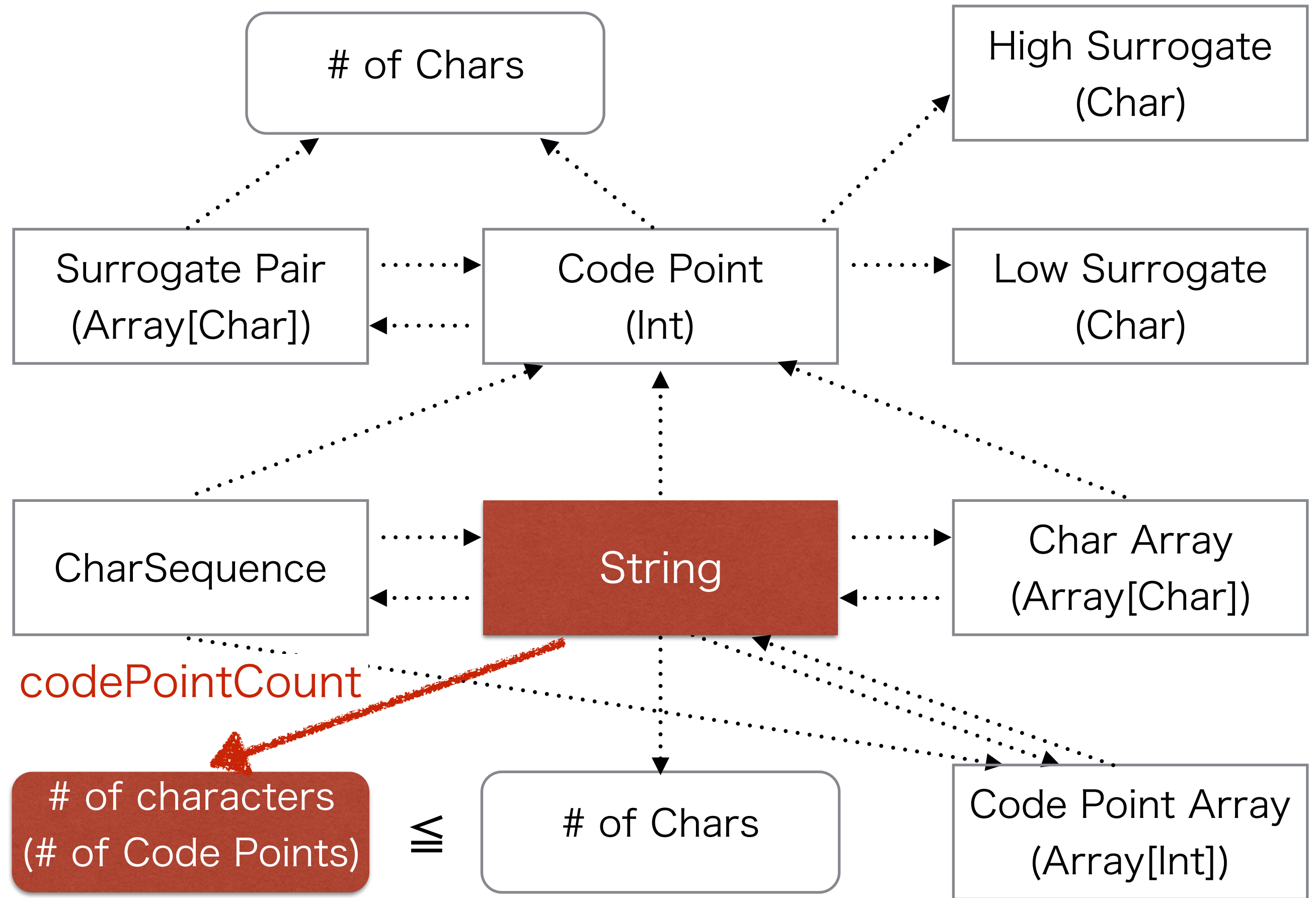
Surrogate Pair が含まれている場合は一致せず。



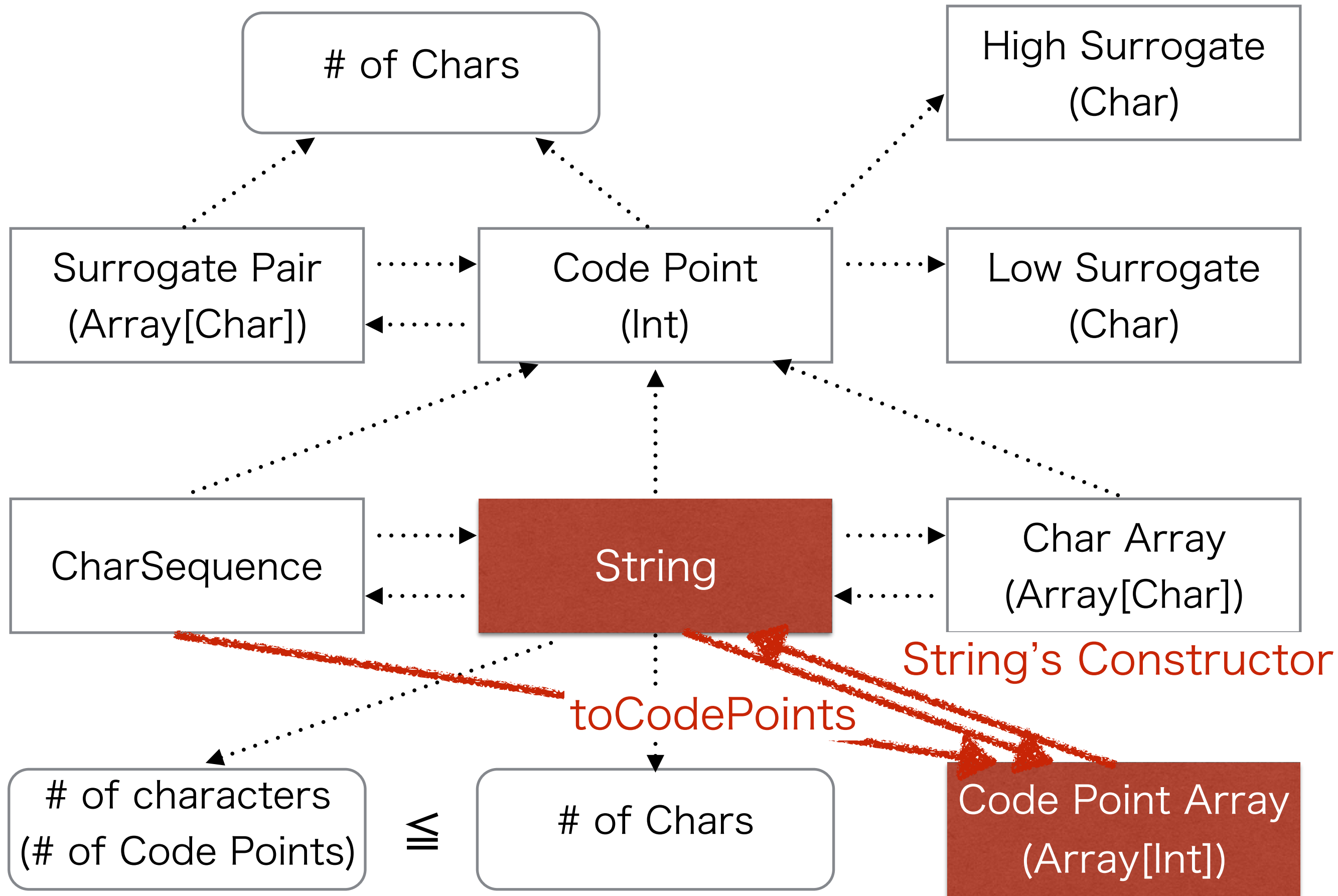
Surrogate Pair が含まれている場合は一致せず。



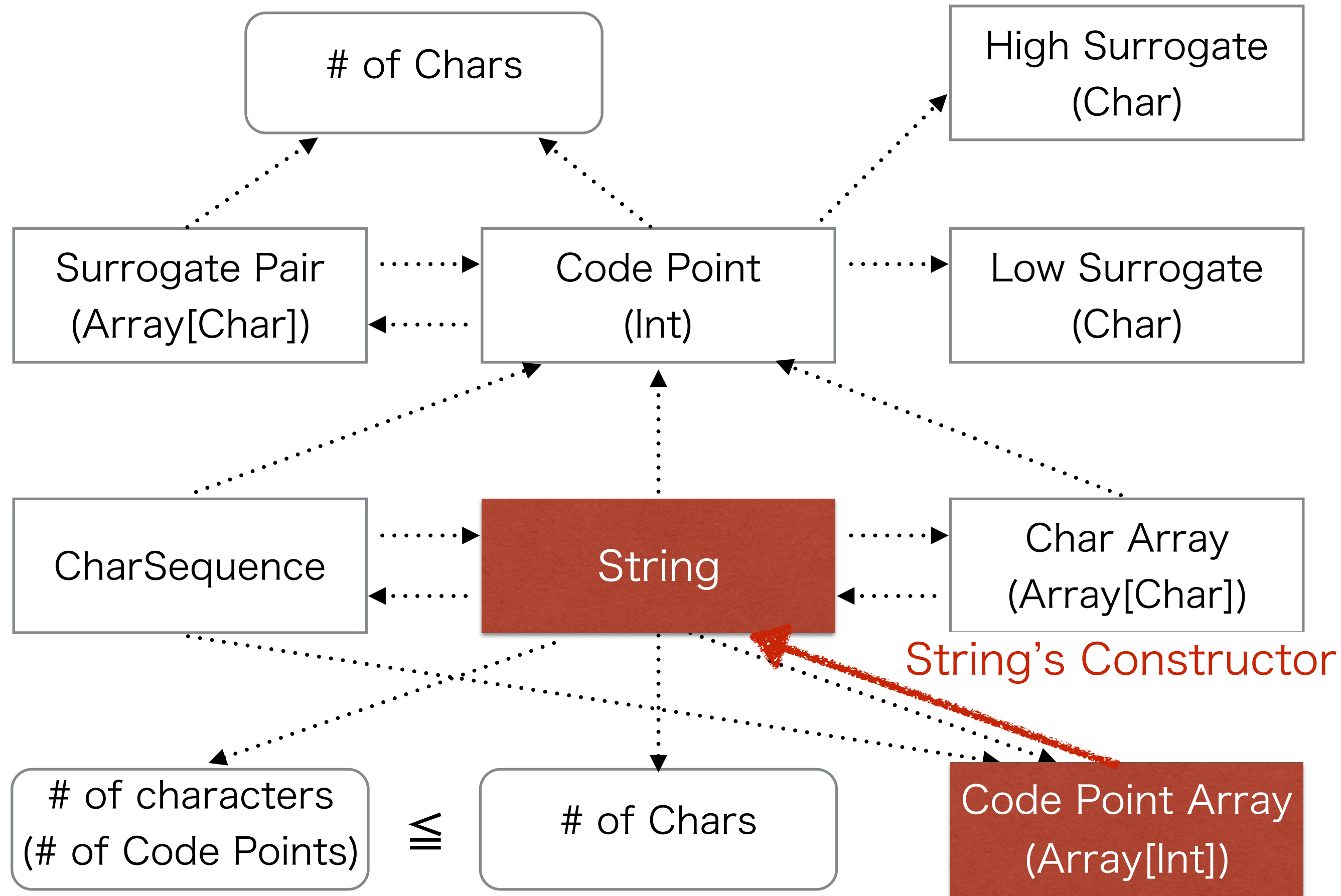
Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。



Surrogate Pair が含まれている場合は一致せず。

コードポイント数だけ移動した 位置のインデックスの取得

`offsetByCodePoints`メソッドは、

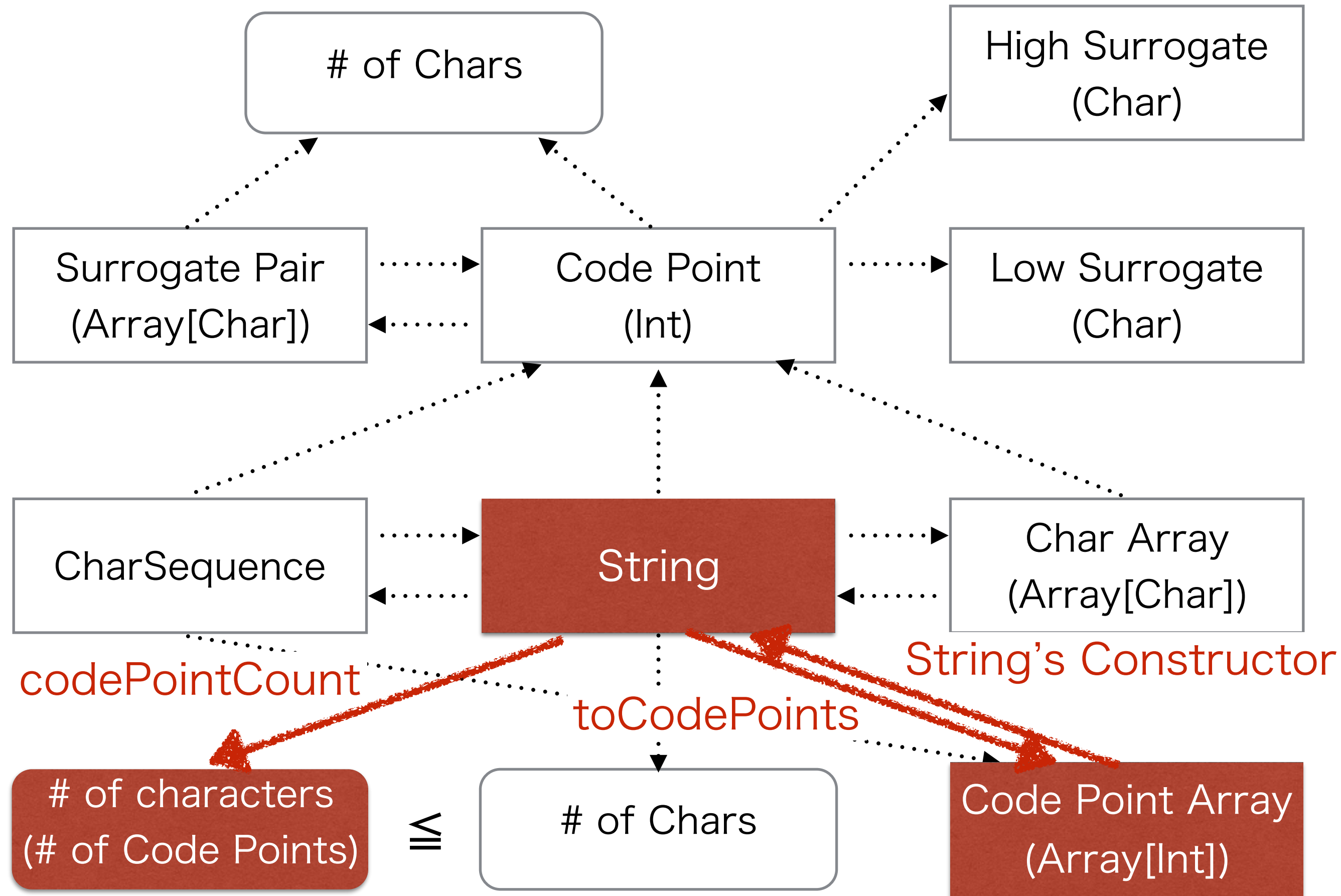
指定された`index`から引数で与えたコードポイント数
だけオフセットされた位置のインデックスを返します。

StringCharacterIterator

Stringを、Char単位でイテレートするCharacterIterator
インターフェースを実装するクラス

CharacterIteratorインターフェース

- ・ firstとnextメソッドで先頭から順方向に回す
- ・ lastとpreviousメソッドで末尾から逆方向に回す
- ・ CharacterIterator.DONEは、CharacterIteratorがテキストの終わりか初めに達したときに返される定数0xFFFF

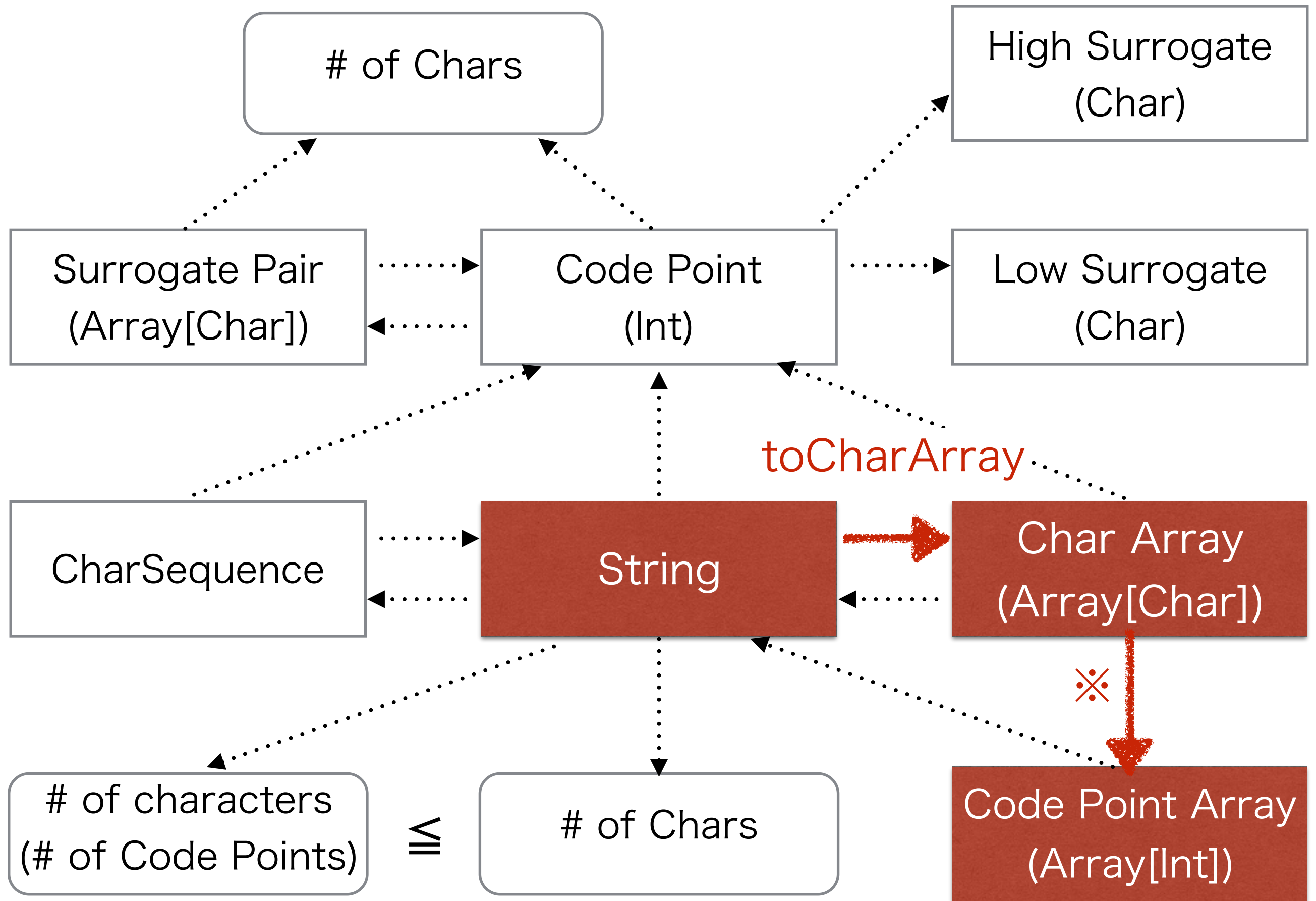


Surrogate Pair が含まれている場合は一致せず。

Java 7以前のStringから コードポイント配列への変換

Java 言語による Unicode サロゲート・プログラミング
(IBMのMasahiko Maederaさんによる技術文書)

- ・ <http://www.ibm.com/developerworks/library/j-unicode/>
- ・ <https://www.ibm.com/developerworks/jp/java/library/j-unicode/>
- ・ https://www.ibm.com/developerworks/jp/ysl/library/java/j-unicode_surrogate/



Surrogate Pair が含まれている場合は一致せず。